AD-A013 583

A NEW TIME-DOMAIN ANALYSIS OF HUMAN SPEECH AND OTHER COMPLEX WAVEFORMS

Janet MacIver Baker

Carnegie-Mellon University

Prepared for:

Air Force Office of Scientific Research
Defense Advanced Research Projects Agency

May 1975

A NEW TIME-DOMAIN ANALYSIS OF HUMAN SPEECH AND OTHER COMPLEX WAVEFORMS

Janet MacIver Baker

May 1975

DEPARTMENT
of
COMPUTER SCIENCE

Carnegie-Mellon University

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR - TR - 75 - 1053 — 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A NEW TIME-DOMAIN ANALYSIS OF HUMAN SPEECH AND OTHER COMPLEX WAVEFORMS	5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Janet MacIver Baker	8. CONTRACT OR GRANT NUMBER(s) F44620-73-C-0074
9. PERFORMING ORGANIZATION NAME AND ADDRESS Carnegie-Mellon University Computer Science Dept. Pittsburgh, PA 15213	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61101D AO-2466
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Blvd Arlington, VA 22209	12. REPORT DATE May 1975 13. NUMBER OF PAGES 158
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Air Force Office of Sciencific Research/NM 1400 Wilson Blvd Arlington, VA 22209	15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The purpose of this research is to explore the usefulness of a new time-domain analysis of complex waveforms, especially with respect to human speech. A chief advantage of time-domain analysis is its precise temporal resolution, which is particularly useful for characterizing very short duration events or regions of rapid change. A significant portion of the speech waveform consists of such regions, especially at phone boundaries. In addition, it is well-known that the intelligibility

Block 20/Abstract

of human speech is in large part, conveyed by acoustic transitional states, generally encompassed by the consonantal elements (c.f. written Hebrew where specific vowel designations are omitted).

In contrast to classical frequency-domain analysis, this time-domain analysis has no inherent bandwidth limitation; that is, there is no trade-off between time and frequency resolution. A major advantage is gained thereby, for certain purposes. On the other hand, acoustic steady states, such as those usually caused by the vowels in human speech, are best characterized in the frequency-domain. Therefore, despite redundancy, much of the information in these two domains is complementary in nature.

The motivation underlying the time-domain studies presented here, arises from compelling evidence, both perceptual and physiological. The perceptual evidence relates to the high intelligibility of infinitely peak-clipped speech, while the physiological evidence relates to the coding operation known as "phase-locking", performed by a large number of first-order auditory neurons. Our signal analyses are directly analagous to these two forms of information extraction. We derive our parameters from the individual waveform cycles, which are defined as occuring between successive up-crossings of the waveform across a zero-axis. A important distinction between this analysis and that of most other zero and up-crossing analyses, is that the cycle measures are not uniformly averaged together; therefore we preserve the precise temporal resolution. In addition, we have developed a visual display, the "Log Inverse Period" or "LIP" plot, which provides a very useful representation of much of this cycle-based information.

Essentially three separate investigations are presented, with the last two predicated on the results of the first.

1) Cycle-based time-domain parameters were extracted from the speech waveforms of many hundreds of utterances, and were then subjected to extensive scrutiny, both by hand and by machine. In addition to investigating a time-domain characterization of speech waveforms in

Continued

general, we found a number of new acoustic phenomena of speech. Many of these are of short duration and/or low amplitude, and are frequently found at phone or sub-phone boundaries. The existence of such events contributes to a more phone-discrete view of continuous speech than is generally held.

2) Based solely on time-domain phenomena found in the previous study, we wrote an automatic segmentation program for continuous speech. Given the same data set and compared against other segmentation programs available in the speech community, the time-domain segmentation (ithout speaker training) compared favorably in all respects, and superiorly in certain respects.

3) We examined the time-domain acoustic characteristics of 228 allophones of fricatives and stop consonants, for each of three speakers (2 males, 1 female). We determined consistent acoustic differences between these, and demonstrated that given this understanding, good discrimination in pair-wise phone comparisons is possible, both among the stop consonants and among the fricatives, even without contextual information.

Finally, we present a personal view of the synergism inherent in the utilization of these time-domain techniques with the traditional frequency-domain techniques. In addition, suggestions are presented for applying these generalizable time-domain techniques to other complex waveforms, especially amenable to such analysis. Specific examples are drawn from music (e.g. violin) and animal (e.g. bou-bou shrike) vocalizations.

A NEW TIME-DOMAIN ANALYSIS OF HUMAN SPEECH AND OTHER COMPLEX WAVEFORMS

Janet MacIver Baker
May 1975

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biocommunication and Computer Science

Mellon Institute of Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania

This work was supported by the Defense Advanced Research Projects Agency under contract F44620-73-C-0074.

ic

Acknowledgements

I especially wish to express my sincerest thanks and gratitude to R. Reddy, for generously providing the support and research freedom necessary to pursue these studies, to Allen Newell, who is largely responsible for the excellent speech research milieu at Carnegie-Mellon University, to J. C. R. Licklider, whose infinite peak-clipping experiments demonstrated the significance of time-domain information, to my long-time mentor, Jerome Y. Lettvin, whose spirit and CLOOGE provided the spark, and to my most generous and understanding colleague, James K. Baker, with whom anything is possible! This research was supported in part by the Advanced Research Projects Agency of the Department of Defense under contract no. F44620-73-C-0074 and monitored by the Air Force Office of Scientific Research. Final editing and revisions of the dissertation were done while the author was with the Speech Processing Group, Computer Science Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

TABLE OF CONTENTS

PREFACE

The central aim of this research investigation is to examine what kinds of information some new techniques of time-domain analysis reveal about the complex waveforms of speech. A chief advantage of time-domain analysis, in general, is its precise temporal resolution, which is particularly useful for studying regions of rapid change as well as very short duration events. A significant portion of the speech waveform consists of such regions, especially at phone boundaries.

The motivation underlying the time-domain studies presented here, arises from compelling evidence, both perceptual and physiological. The perceptual evidence relates to the high intelligibility of infinitely peak-clipped speech, while the physiological evidence relates to the coding operation known as "phase-locking", performed by a large number of the first-order auditory neurons. Infinite peak-clipping rectifies the waveform (optionally with or without preservation of peak and valley amplitude information), thereby preserving only the times of waveform up-crossings and down-crossings relative to a horizontal zero-axis drawn through the waveform, along with the appropriate polarity designation. That is, as the waveform crosses the axis, the information retained consists of the time of the crossing and its direction, up or down. In contrast, given a signal waveform, a phase-locking neuron fires once, phase-consistently, for each cycle or integer multiple number of cycles within the waveform. Essentially, the operation of phase-locking is equivalent to infinite peak-clipping which responds only to either waveform up-crossings or down-crossings, but not both(the information contained in a set of crossings of either polarity is highly redundant with that contained in the set of opposite polarity). Our signal analyses are directly analogous to these two forms of information extraction. We derive our parameters from the individual cycles in a waveform. A cycle is defined as that portion of waveform occuring between two consecutive up-crossings of the waveform across a zero-axis drawn through its center. The inverse of the cycle period (the duration of time between successive up-crossings) is refered to as "cycle-frequency" The other parameters used are also extracted from individual cycles of the waveform. In the visual display we have developed, the logarithm of the fundamental inverse period or cycle-frequency measure is plotted on the y-axis against time on the x-axis. This display is named the "Log Inverse Period" or "**LIP**" plot.

In the exploration of of our time-domain methods, we have conducted both basic and applied research. Presented here are three separate investigations; the last two are predicated on the results of the first.

1) Cycle-based time-domain parameters were extracted from speech waveforms, and then subjected to extensive scrutiny, both by hand and machine. Many hundreds of utterances were analyzed. These included sizable numbers each of nonsense syllables, citation form phrases, and complete sentences of connected speech, with all forms of speech spoken by both males and females. These studies have enabled us to describe in the time-domain many of the acoustic characteristics of speech well-known from traditional frequency-domain studies, as well as to discover many new acoustic phenomena of speech. Some of these phenomena comprise very short duration events, often low in amplitude,which occur at phone boundaries or even sub-phone boundaries where acoustic states change rapidly. The existence of such events contributes to a more phone-discrete view of continuous speech than is generally held.

2) Based solely on the time-domain phenomena found and investigated in the previous study, we wrote a program for automatic segmentation of continuous speech. This program was run on a set of utterances submitted to the Segmentation and Labeling Workshop held at Carnegie-Mellon University (July,1973). The results were then compared with those presented by other groups of the speech community. This evaluation demonstrated that the time-domain based segmentation (without speaker training) compared favorably in all respects, and superiorly in certain respects, to the segmentation results of the other programs.

3) We studied differences in the time-domain acoustic characteristics among a set of 228 allophones of fricatives and stop consonants, spoken by each of three speakers (2 male, 1 female). The aim of this study was to examine if different commonly occurring contexts of a given phone cause changes in the acoustic manifestations of that phone. And if such changes do occur, are they regular? For example, is a /p/ in a retroflexed environment acoustically different from a /p/ in a nasalized environment? The answer is "yes". What's more, the effect of retreflexion on a /p/ is

similar to that on a /b/,/d/,/t/,/g/, and /k/! We performed stringent pair-wise phone recognition tests which were based on the computed allophone effects derived from single instances of phones not included in the test set. For example, we tried to distinguish /k/ allophones from /t/ allophones, based on the allophone statistics collected from the combined set of allophones from /b/,/p/,/g/, and /d/ by the same speaker. Significant statistical results for phone recognition were obtained, thereby supporting the concept of regular acoustic allophone transformations for commonly occurring fricatives and stop consonants of general English.

Finally, we present a personal view of the synergism inherent in the utilization of these time-domain techniques with the traditional frequency-domain techniques. In addition, suggestions are presented for applying these generalizable time-domain methods to other complex waveforms, especially amenable to such analysis. Specific examples are drawn from music and animal vocalizations

Chap. er I - THE TIME-DOMAIN AND SPEECH

INTRODUCTION

The acoustics of human speech a· ¹ of other complex animal vocalizations has long been an area of primary interest to a number of researchers. Basic findings in this area are directly relevant for linguistic studies, speech aids for the handicapped, child speech development, cross-cultural language speech comparisons, speech compression, automatic speech recognition, physiological processes, perception, and so forth.

THE NEED FOR TIME-DOMAIN INFORMATION

Over the years, tremendous amounts of time, energy, and expense have bee. devoted to studying the acoustics of speech, both for basic research and for applications such as those listed above. These investigations have chiefly centered around analyses of speech waveforms, both analog and digital. These speech waveforms are usually derived from the voltage-time relations directly proportional to changes in air pressure caused by speech. A speech waveform plot, referred to as an "oscillogram", contains complete information about the original signal from which it is derived. An oscillogram is an example of a time-domain display because time is determined for any given event or feature occurring in the signal; for example, the time of maximum amplitude occurrence during a pitch period. Usually the speech waveform itself exhibits a great deal of variability in a continuous fashion. In general, except for stressed vowels and central portions of slowly spoken phones, a given acoustic state .oon transitions into a different one. Unfortunately, past studies of time-domain methods have not revealed good ways of reducing this large bulk of highly variable data and deriving from it robust and useful parameters for speech analysis. Therefore time-domain analyses of the waveform have generally been relegated a very minor role in speech studies, the major role being played by frequency-domain analyses.

Frequency-domain analyses, largely dominated by spectrographic studies, average waveform frequency components over some fixed interval of time in order to derive spectral measures of the signal. This averaging means that in the frequency-domain, there exists an inherent bandwidth limitation such that the better the frequency resolution, the worse the time resolution, and vice versa. For steady state periodic or quasi-periodic signals, frequency-domain analyses reveal good information on the waveform component frequencies. In human speech, chiefly vowels satisfy this criteria and are therefore most amenable to this analysis. Vowels, especially stressed vowels, are characterized by waveforms which are relatively steady state with pitch periods consisting of several cycles each. Frequency-domain methods, with speaker training, have proven quite successful in reliably characterizing vowels, especially with respect to formant structures.However, many other phones with non-periodic waveforms, often undergoing rapid frequency and amplitude modulations, are much more poorly described in the frequency-domain. Here the bandwidth limitation becomes a serious hindrance because the time during which an acoustic feature of a consonant or phone change occurs, is often too short for good frequency discriminations. The implications of frequency-averaging in such cases must be carefully considered.

When classical frequency-domain methods are applied to a signal composed of rapidly changing frequency characteristics, the result is an average of its component frequencies, which does not bear a unique relationship to the original signal. Clearly, averaging over such a period of changing frequency does not yield a good representation of the original signal. It is also true that the magnitude spectrum of a given waveform over the interval of averaging is the same whether the waveform is played frontwards,backwards, or otherwise transformed in various ways.

The bandwidth limitations of frequency-domain methods are not serious providing these hypothetical fast-changing signals do not occur or occur infrequently. However, even simple visual inspection of speech waveforms reveals an abundance of such situations.

These situations frequently occur at phonetic boundaries and during the course of stops and fricatives, all regions characterized by acoustic transients. Therefore, due to the acoustic structure of the waveform itself, some of these features are only observable with precise temporal resolution of the signal. Therefore, time-domain techniques capable of such resolution, are required.

In addition, it is well-known that the intelligibility of human speech is, in large part, conveyed by acoustic transitional states, generally encompassed by the consonantal elements (cf. written Hebrew, where specific vowel designations are omitted). The vowels, acoustically relatively steady state, although well described and discriminated on the basis of spectrographic features, assume a lesser role in speech intelligibility. Nonetheless, both the acoustic features observed in the time-domain and those observed in the frequency-domain are important for speech characterization. That is, analyses in both domains are complementary. In addition, due to the well-known redundancy in speech, analysis in the time-domain yields information found by frequency-domain studies, and vice versa. That is, time-domain analyses reveal some information about the steady state vowels, just as frequency-domain analyses reveal some information about the transitional state consonants.

HOW TO FIND TIME-DOMAIN INFORMATION

Having made a case for the necessity of characterizing acoustic transients in the time-domain, we therefore propose and demonstrate a means for doing so. We have extensively investigated a kind of time-domain analysis which is characterized by parameterizations containing much of the waveform information necessary for the intelligibility of speech. These parameters are computed for each individual cycle in the speech waveform. A cycle is designated as that portion of waveform occurring between successive up-crossings of the waveform across a zero-axis drawn horizontally through the center of the waveform. Three kinds of cycle measures have proven very useful: measures of the period or cycle duration, amplitude, and cycle microstructure. Of these, the most interesting is the cycle period which measures the duration of a cycle between successive up-crossings. Hereafter, the inverse or reciprocal of this period, is referred to as "cycle-frequency". The choice of this particular parameter is motivated by the compelling evidence derived from perceptual studies on infinitely peak-clipped speech and on the neurophysiological studies of phase-locking in auditory information information processing.

The perceptual studies were conducted about 25 years ago by Licklider and his colleagues [L1,L2], who demonstrated the high intelligibility of infinitely peak-clipped speech. Infinite peak-clipping reduces the complex waveform to a rectangular waveform where all cycles are equal in amplitude. The only information retained by this transformation is the time and direction of waveform zero-crossings. These experiments clearly showed that a great deal of information must be encoded in the temporal pattern of zero-crossings of the speech waveform alone! Since the time-domain parameters which completely characterize this information are also easy to obtain automatically, we have thoroughly examined these and other related time-domain parameters in order to discover what they may reveal about speech. Since speech is very redundant however, this information may well be encoded in other acoustic features of the waveform as well.

The idea of looking at zero-crossing measures *per se* is not in itself conceptually new; following Licklider's studies, other investigators [C2,C3,C4,S1] have looked at zero-crossings and up-crossings. However, in contrast to most of these other investigators [I1,M3,R1] who have used zero-crossings to analyze speech, we do not average the up-crossings over a fixed interval of time. Reasons for this will be discussed shortly. First of all, it is important to be aware that the chief motivation for most up-crossing studies has been in searching for an inexpensive way to find frequency-domain acoustic features such as formants [D1,M4,P1] The zero-crossing methods avoided the computations required for Fourier transforms, for example. In order to decrease the expense and variability in examining individual cycles (and in fact no previous claim had been made attaching any significance to individual cycles), it was easy to compute an average cycle length by simply counting the number of zero-crossings occurring during a given time interval. This procedure has two major consequences: 1) the perfect time resolution inherent in the time-domain is lost when the crossings are averaged; that is, a bandwidth limitation is introduced, 2) the conventional acoustic features extracted are usually less precise and more variable than the same features extracted directly in the frequency-domain. More recently, the techniques of linear prediction have been popularized [A1,M1,M2]. Linear prediction consists of finding the coefficients for a linear filter which minimizes the least squared prediction error, averaged over a given analysis interval. The non-stationary formulation of linear prediction is a time-domain technique [C1]. However, the analysis interval or window used (though not inherently determined by the technique itself) generally consists of a 10 msec duration or a pitch period. Therefore, here also,

Our reason for not averaging up-crossings generally, is that in the speech waveform itself there are significant acoustic features which last for only one or a few cycles in duration. If such cycles are averaged in with others, this information is irrevocably lost. As previously mentioned, such transient events frequently occur at phone boundaries as well as between other acoustically distinct regions, within stop consonants, for example. We do consider it appropriate however, to average our time-domain parameters across regions which are acoustically uniform, such as the frication region of a fricative, exclusive of any boundary transition events. Therefore, when we refer to an "average" parameter value, only the parameters of cycles within such a uniform acoustic region are averaged together.

The second motivation for this work comes from neurophysiological research on the auditory information processing of the ear itself [D1,F1,G1,K2,K3,K4]. The information conveyed in an incoming signal is encoded by different kinds of neurons in the ear and then transformed and integrated at higher neuronal levels. Each of the neurons communicates information in a very simple form; namely, by propagating a sequence of electrical impulses or spikes, each of which is the same voltage or amplitude. This one dimensional response is referred to as "all-or-none". Currently, the general consensus of auditory neurophysiologists recognizes that the ear codes different aspects of an auditory signal both spectrally and temporally; that is, in both the frequency-domain and the time-domain, respectively. The frequency-domain analysis performed by the ear is analogous to that performed by a filter bank. Different neurons along the basilar membrane respond to different frequency ranges; that is, a neuron fires if it detects a signal of sufficient intensity within a given frequency range. Neurons also code information in the time-domain, in a fashion known as "phase-locking"[K1,R2]. Given a representation of the waveform, a phase-locking neuron responds by firing once, phase consistently, for each cycle or integer multiple number of cycles within the waveform. Phase-locking occurs for signal frequencies, from the lowest audible frequencies up to at least 4.5 kHz and perhaps above 7 kHz. This frequency range encompasses the chief information bearing portion of the speech spectrum. It is very likely that the ear has evolved such that both the time-domain and frequency-domain information derived from acoustic signals is integrated and utilized synergistically.

The analogy between infinite peak-clipping and phase-locking should now become clear. Both are very similar time-domain transformations of the acoustic signal except that the former method temporally characterizes two phase consistent aspects of the waveform whereas the latter characterizes only one. We have previously noted that the kinds of information obtained from up-crossings and down-crossings for any given signal are highly redundant. The time-domain techniques we have developed, directly examine the information available from infinite peak-clipping and phase-locking. In addition, we have examined the usefulness of certain other time-domain parameters derived on a cycle-by-cycle basis.

The time-domain parameters which we have found to be useful are described for a cycle where t1 is its initial up-crossing, t2 is its down-crossing, and t3 is its end or the time of the next waveform up-crossing. The data is sampled at 20 kHz. Therefore the accuracy of the individual sample is 50 microseconds. In order to determine that a waveform up-crossing has occurred, the cycle must have already achieved a negative amplitude value less than minus epsilon (where epsilon generally has been chosen to be equal to 3, out of a range 0-255) and where the most recent sample has a positive value greater than positive epsilon. Then a linear interpolation is performed between this sample and the last negative sample in order to more accurately ascertain the true time of up-crossing. The period (P) of a cycle equals

(1) $P = t3 - t1$

and the cycle-frequency (CF) equals

(2) $CF = 1/P$.

Peak amplitude for a cycle is simply the maximum positive amplitude observed, Amax, during that cycle. Similarly, valley amplitude for a cycle is the most negative amplitude observed, Amin, during that cycle. Absolute amplitude (Absamp) equals

(3) $Absamp = Amax - Amin$.

The parameters total variation and microstructure are both measures of cycle "smoothness". These indicate the presence of higher frequency components of sufficient amplitude to "ride" on a lower frequency "carrier" cycle without causing up-crossings of its own (except near the zero-crossings of the carrier cycle). The total variation (TV) equals:

(4) $TV = \Sigma_{t=1,n} | a(t) - a(t-1) | \; / \; \Sigma_{t=1,n} | a(t) |$ where n = # samples in that cycle.

The microstructure (MS) equals

(5) $MS = (\Sigma_{t=1,n} | a(t) - a(t-1) | - 2*Absamp) / \Sigma_{t=1,n} | a(t) |$

where it should be noted that 2*Absamp is the un-normalized TV of a sine, triangle or square wave

(un-normalized TV = $\Sigma_{t=1,n} | a(t) - a(t-1) |$).

For example in the time-domain, these latter two features provide a major distinguishing feature between the high front vowel /i/ and the high back vowel /u/. The phone /i/ is generally characterized by much higher values of TV and MS than is /u/.

We have developed a display which visually captures the information contained in either the up-crossings or down-crossings of infinitely peak-clipped speech as well as the information conveyed by a neuron phase locked to the speech waveform. We call this display a "Log Inverse Period" or "LIP" plot. Originally this display was suggested by Lettvin who has designed an analog circuit, the CLOOGE (Continuing Log Of On-Going Events), which detects both waveform up-crossings and cycle amplitude, and then displays these, in real time, on an oscilloscope screen which may be continuously photographed. This display is derived from the "instantaneous frequency" plots of single unit activity, used by neurophysiologists.

The initial investigations of these time-domain procedures, using this real-time analog hardware, were conducted by the author in collaboration with Lettvin in his laboratory at M.I.T. However, the LIP plots and most of the time-domain information presented here, are the results of speech data digitally sampled at 20kHz and processed on a PDP-10 in the Computer Science Department at Carnegie-Mellon University. At the time of recording, the signal was band pass

1: The author was a research affiliate (1969-72) of the Research Laboratory of Electronics, M.I.T., and worked with J. Y. Lettvin (professor in the departments of electrical engineering and biology).

filtered between about 100 Hz (for attenuation of 60 Hz electrical hum) and about 8 kHz (to prevent aliasing above the Nyquist frequency of 10 kHz, given a sampling frequency of 20 kHz).

We generate our visual display as follows: a zero-axis is drawn horizontally through the acoustic waveform. We note the exact time when the waveform crosses this axis in an upward direction. Only those up-crossings are recorded which, following a negative excursion of the waveform, then exceed an amplitude threshold, epsilon, set slightly above the zero-axis. This threshold tends to preclude very low amplitude background noise. We measure each interval between successive up-crossings and plot these as a function of time in our displays. Therefore each up-crossing in the acoustic waveform is represented by a discrete dot. We plot, on a logarithmic scale, the inverse of the interval between successive up-crossings; that is, the reciprocal of the cycle period, along the y-axis, and time along the x-axis. This yields a display which superficially resembles a kind of spectrographic display. We also display a rough intensity measure by means of a z-axis modulation. The size of a dot representing a given cycle is proportionate to the log of the greatest amplitude observed during that cycle. This dot size intensity measure of our LIP displays is analogous to the intensity measure expressed in spectrograms. The following illustration (Fig. 1) shows the relationship of the log inverse period plot to the waveform from which it is generated. Note that individual cycle-frequency values may be easily read from the y-axis. In the LIP plots which are used to illustrate the following chapters, the time-scale has been compressed more than in this graphical illustration of the LIP plot. Therefore, even though dots may overlap or appear to occur simultaneously, they actually are occurring at discrete times.

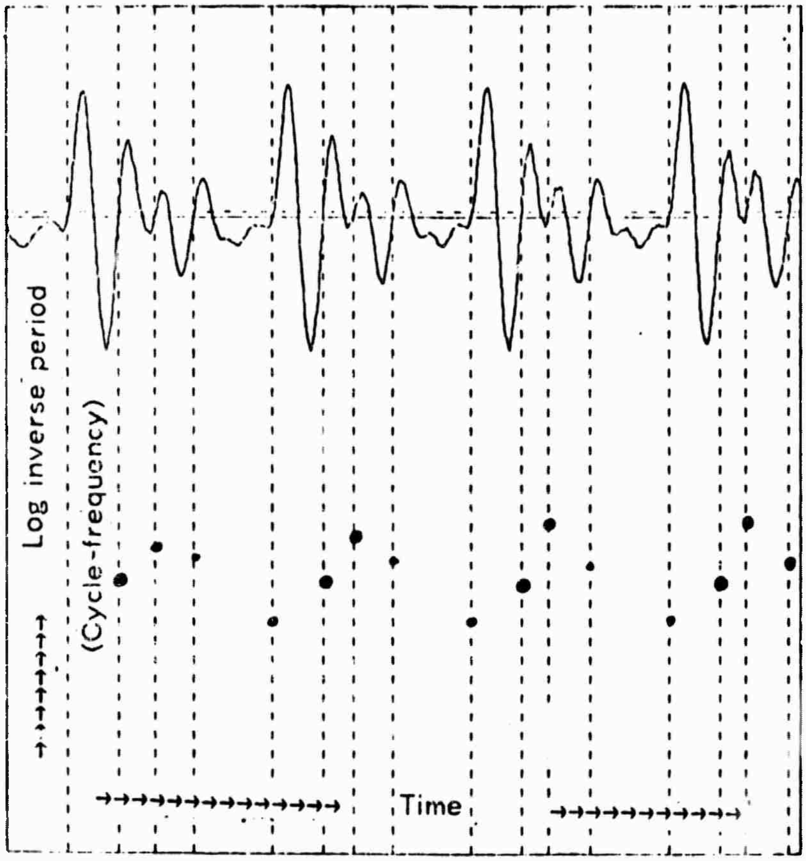


Figure 1: The Log Inverse Period Plot

Chapter II - CHARACTERISTICS OF SPEECH IN THE TIME-DOMAIN

INTRODUCTION

The greatest potential value for time-domain analysis of speech rests with how well it can detect and characterize acoustic transients. As previously described, these are principally found at certain phone boundaries as well as at other acoustically distinct regions, such as within stop consonants, for example. For this reason, we have chosen to concentrate on the characteristics of fricatives and stop consonants.

This is a large undertaking not only because the time-domain analyses we are using are radically different from the classical frequency-domain techniques, but also because we are searching through a much larger data base than is customarily used in order 1) to identify reliably relevant elemental time-domain characteristics of speech, and 2) to ascertain the significance of these. The total amount of data examined in detail during the course of these studies consisted of many hendreds of utterances derived from large sets of nonsense syllables, citation-form phrases, and complete sentences of connected speech, spoken by more than 20 male and female speakers, including adults and children, often in noisy environments. The subset of this data which has been studied most thoroughly, consists of 684 phrases in citation form, generously provided by J. Shoup. Each of the speakers (2 males, 1 female) spoke 228 utterances chosen to provide examples of most of the allophones of the fricatives and stop consonants, including those common in general English as described by Shoup [S3]. Further descriptions of this set of data will be provided in Chapter IV which discusses allophonic differences in the fricatives and stop consonants. In searching for meaningful aspects of the acoustic signal encoded by time-domain parameters, we have been guided by an understanding of the primary principles of single-unit stimulus-response characteristics in the nervous system. Operationally this has meant careful study of acoustic regions where sharp discontinuities consistently occur along one or more dimensions.

The present chapter discusses the more ubiquitous acoustic phenomena revealed by our time-domain analyses. First is a discussion of the canonical forms of fricatives and stop consonants, next a description of some related acoustic-phonological phenomena, and finally some observations on time-domain characteristics of the other phones. Both LIP plots and waveforms are provided to illustrate each of the phenomena discussed. It is highly recommended that the reader spend some time carefully examining these in order to understand the relationships of information representation in each of these displays. The reader is advised to remember that the LIP plot is a visual representation of precise cycle-frequency information along with some amplitude information. However, in the description of the acoustic features of the phones, characteristic relative changes in cycle-frequency, amplitude, and microstructure measures will be stated as well.

Each line of waveform has a duration of .1 sec. These lines read left-to-right, from top-to-bottom, are consecutive. The beginning and end times of each complete waveform are designated by arrows on the time-axis of the corresponding LIP plot. On some LIP plots, vertical lines are drawn as segmentation boundary markers between different acoustic states. In thse plots, phone labels for the different segments are printed along the time axis. Since there exists some background noise for most of the speech shown here, waveform up-crossings normally occur during intervals of "silence". The LIP dots representing this noise are generally small, reflecting the low amplitude of such cycles. During intervals of speech, this background noise is superceded by the greater amplitude of the speech signal. In order to give the reader a more concrete concept of the parameters measured, single examples of characteristically typical phones will be described with some quantitative details. In addition, the frequency of occurrence for certain time-domain observed features of speech will be given at the conclusion of this chapter.

FRICATIVES

The set of fricatives studied consists of /v/, /f/, /ð/, /θ/, /z/, /s/, /ʒ/, and /ʃ/. Generally the fricatives are acoustically characterized by sustained high frequency regions. In voiced fricatives, this high frequency region is preceded by a low frequency region, the familiar voice-bar, which may persist throughout the high frequency region as well. Time-domain analysis reveals that at the beginning of the high frequency region of the fricative, there are very sharp discontinuities occurring simultaneously, upward for both cycle-frequency and microstructure, and often a sharp decrease in amplitude where the fricative is preceded by a vowel. The new acoustic state which results from these large changes, is usually sustained for most of the fricative duration. Usually at the end of the fricative, sharp discontinuities with respect to cycle-frequency, amplitude, and microstructure, are again observed, at the boundary between the fricative and the following phone. However, a different transient kind of acoustic event often occurs at the very beginning and again at the very end of the fricative. Sometimes occurring at these places is one or a few cycles characterized by lower cycle-frequencies than those of the other cycles in the acoustic segment immediately preceding and in the acoustic segment immediately following this transitional phenomenon. Amplitude of these cycles is variable, although the cycle microstructure is usually low. These transition cycles are marked "t" in the LIP plots. Regions of frication are marked "f", and for voiced fricatives, the initial voicing region is marked "v". The first example (Fig. 2) is an /s/ from the utterance "there sir" (HN,♀). The duration of the frication region is .17 sec. The cycle-frequency of the transition cycle into the frication is 201 Hz and the cycle-frequency of the transition cycle at the end of the frication is 304 Hz. Within a few cycles of the initial transition cycle, the average amplitude of the cycles in the fricated region drops to a fraction of its value prior to the transition, while the cycle-frequency and microstructure measures sharply increase. The same kinds of changes occur, although in the opposite direction, as the frication abruptly ends at the final transition cycle, and the following vowel commences.

The second example (Fig. 3) shows the voiced fricative /v/ in the utterance "invent" (EH,♂). In this utterance, there occurs first a transition cycle with a cycle-frequency of 154 Hz followed by four cycles of voicing (average cycle-frequency for this acoustic region is 206 Hz), which precede the /v/ frication. This frication lasts for 42.3 msec and is terminated by two end transition cycles with cycle-frequencies of 375 Hz and 361 Hz, respectively. Here the initial sharp discontinuities, upward for cycle-frequency and microstructure, and downward for amplitude, commence not with the initial voicing portion of the /v/ but with the onset of frication. This is typical of a voiced fricative. And here too, sharp discontinuities in the opposite direction occur within a few cycles of the final transition cycle.

Example representations follow for each of the remaining fricatives. The phones and utterances are /ð/ (Fig. 4) in "Display the phonemic" (LM,♀), /θ/ (Fig. 5) in "to thaw" (EH,♂), /z/ (Fig. 6) in "hers earns" (HN,♀), /f/ (Fig. 7) in "the feed" (JA,♂), /ʒ/ (Fig. 8) in "rouge was" (HN♀), and /ʃ/ (Fig. 9) in "her shirt" (EH,♂).

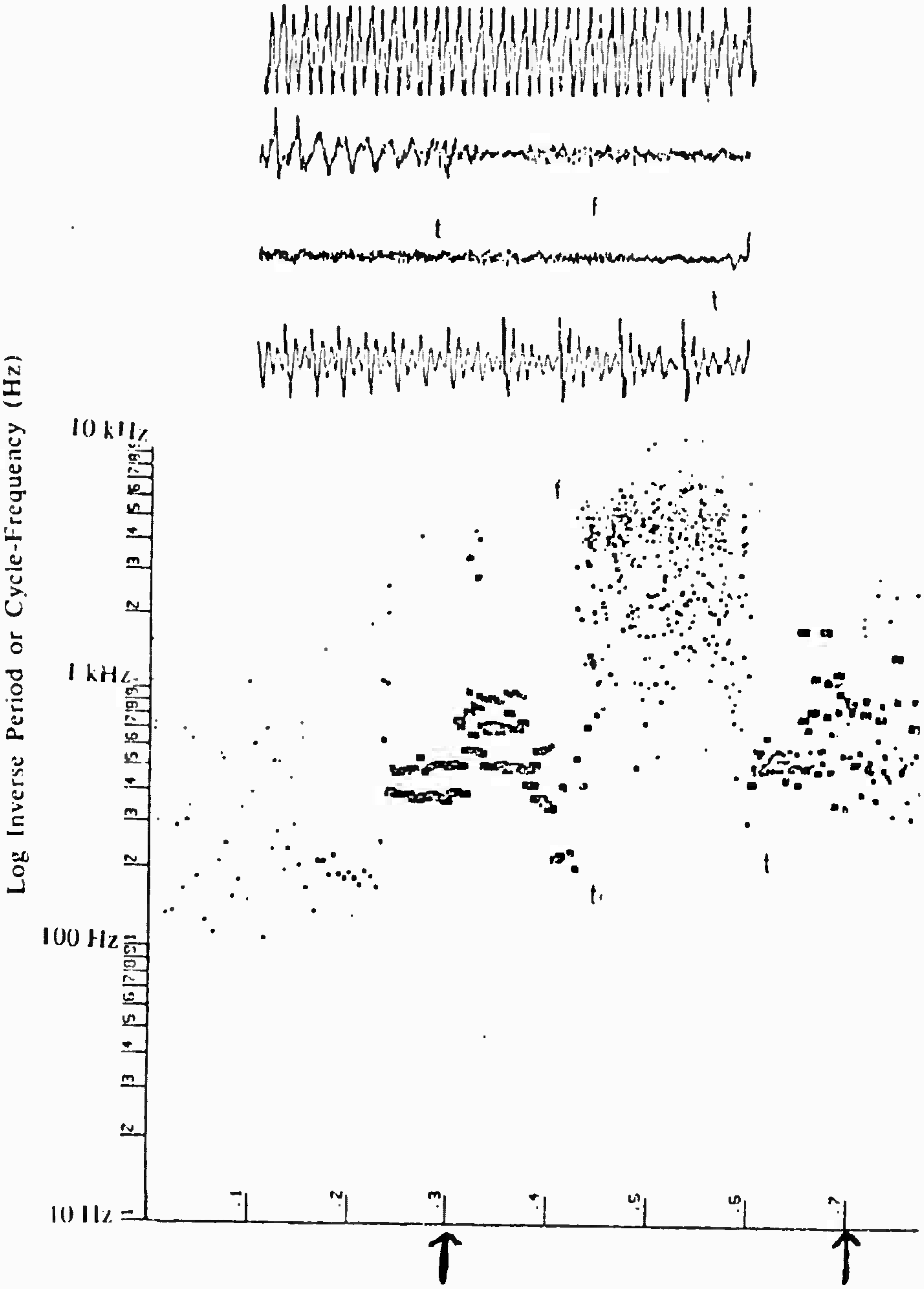


Figure 2: there sir (♀,HN)

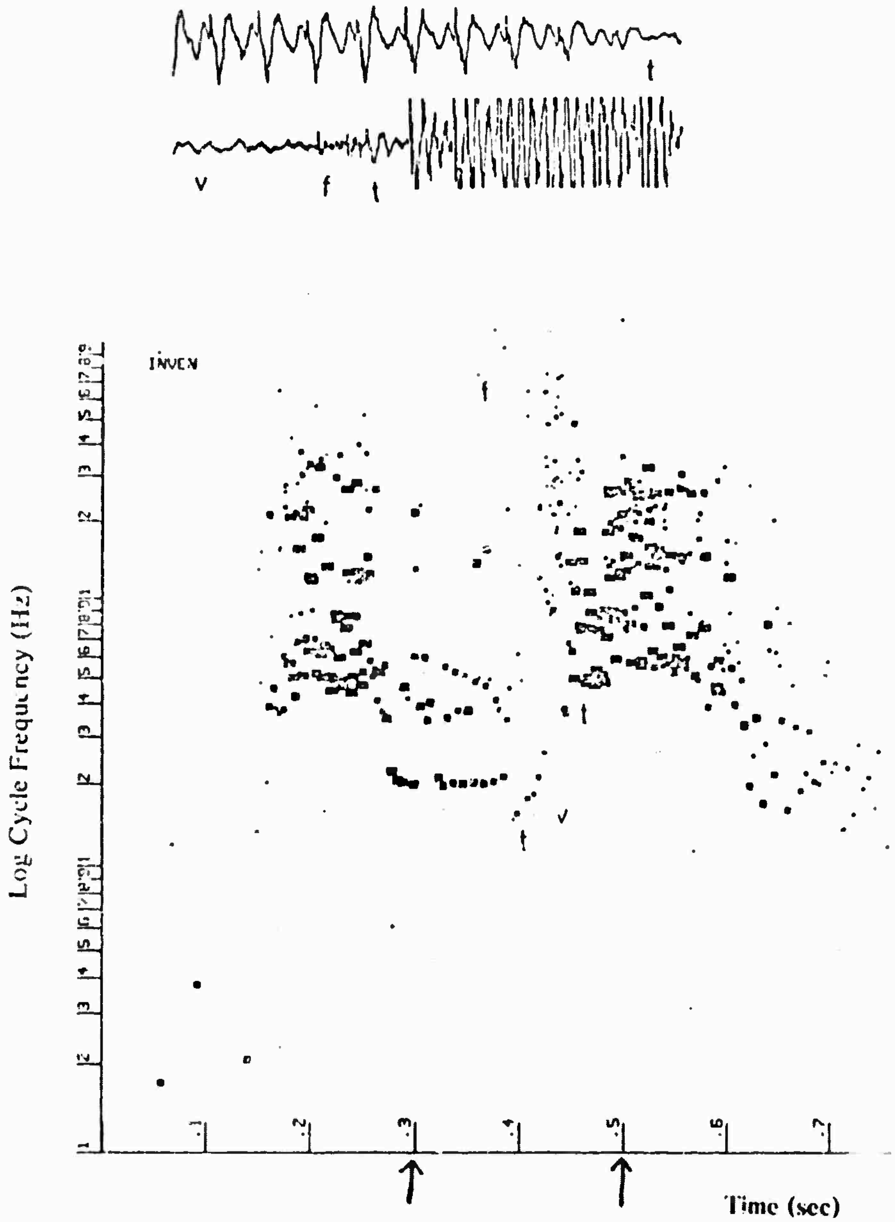


Figure 3: invent(σ,EH)

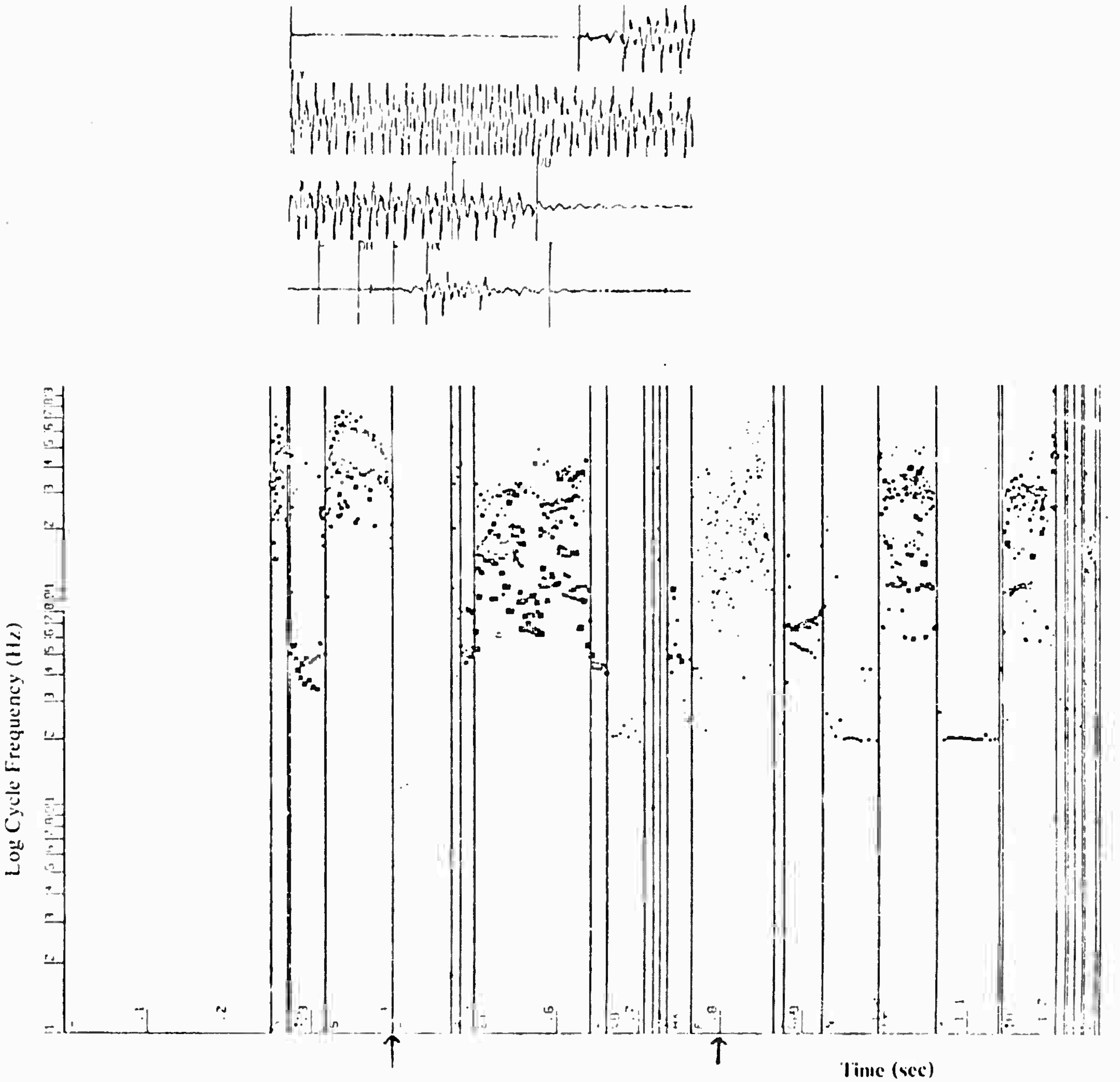


Figure 4: Display the phonemic(ọ,LM)

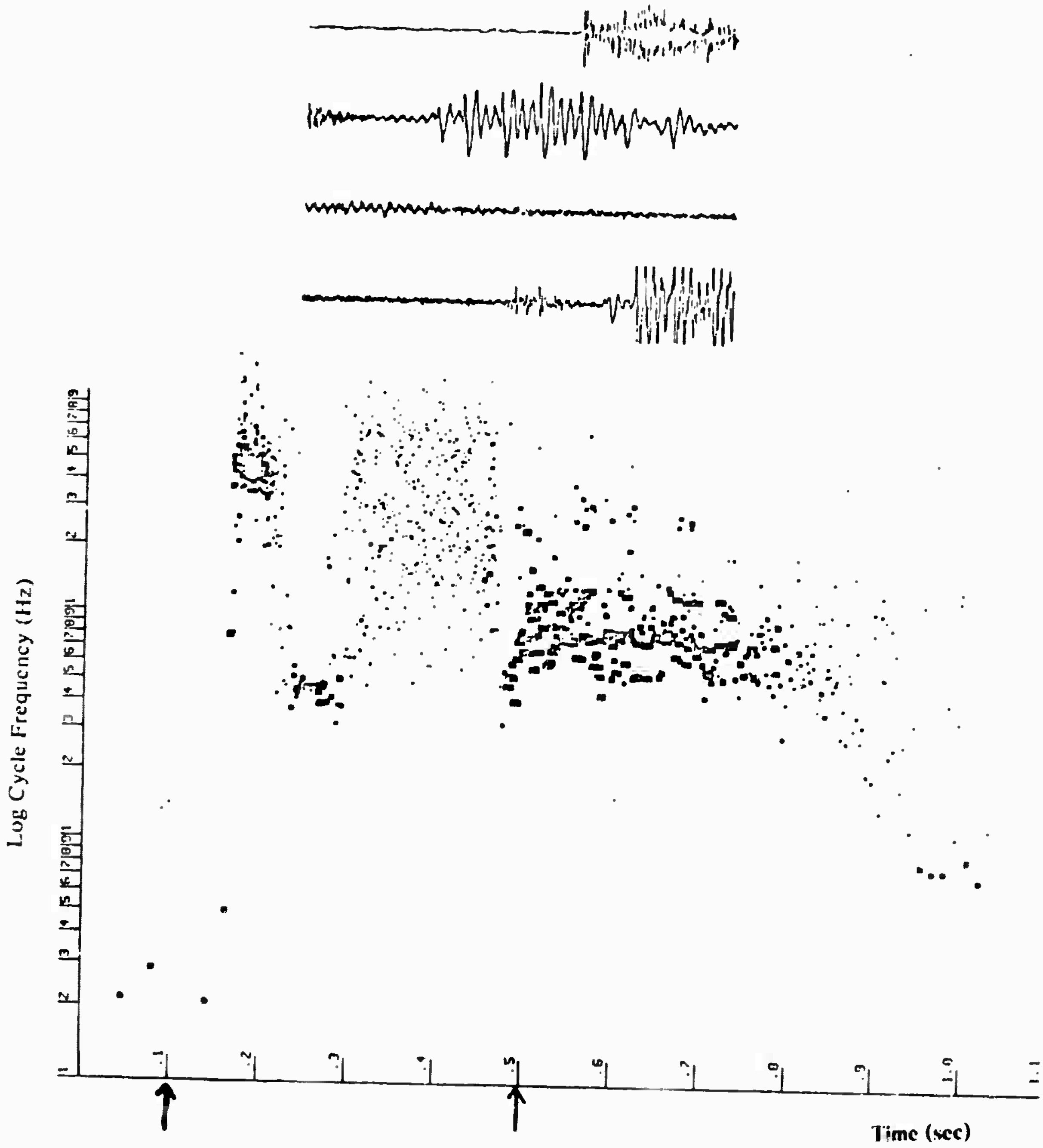


Figure 5: to thaw (σ,EH)

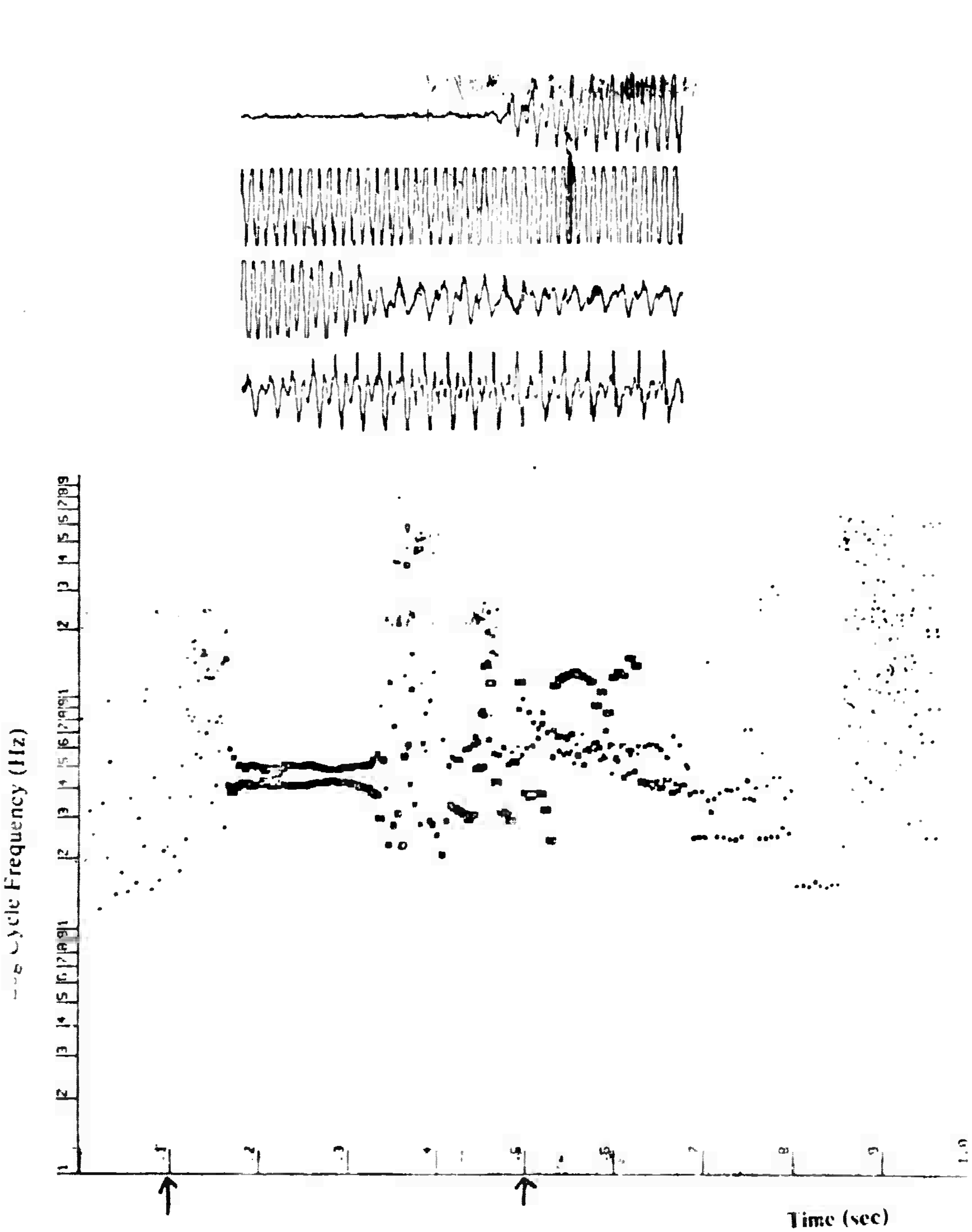


Figure 6: hers earns (♀,HN)

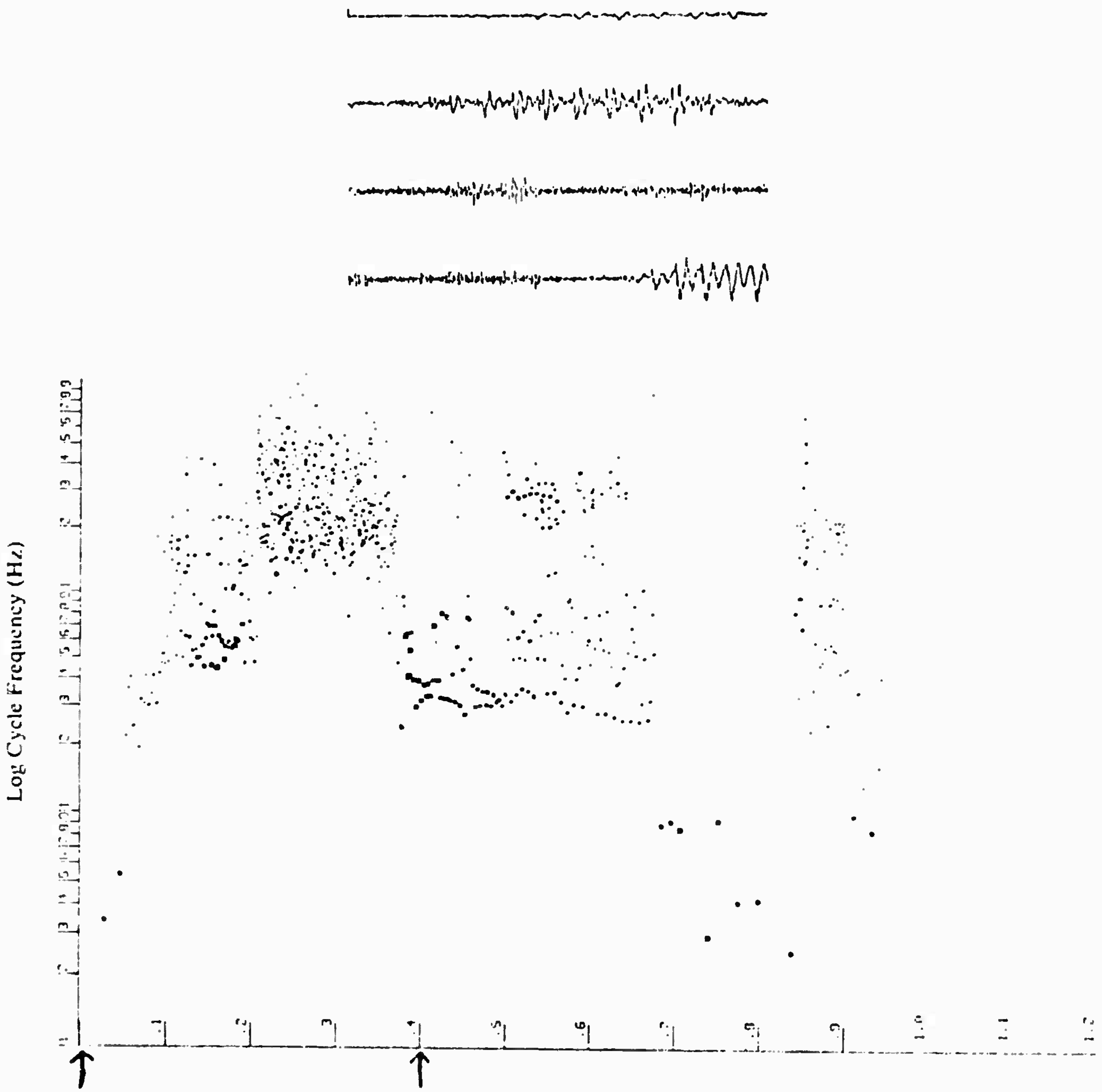


Figure 7: the feed (σ,JA)

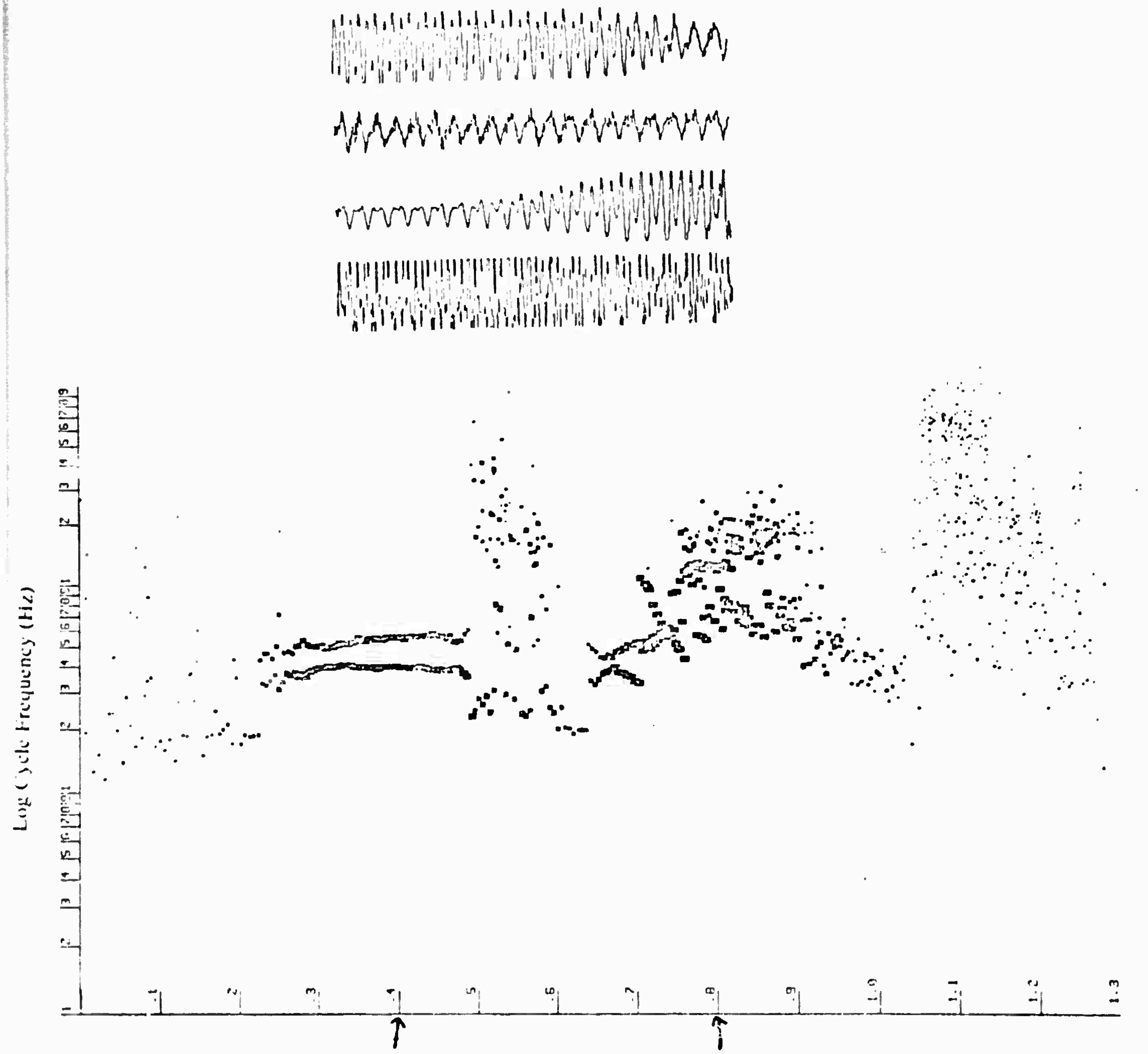


Figure 8: rouge was (♀,HN)

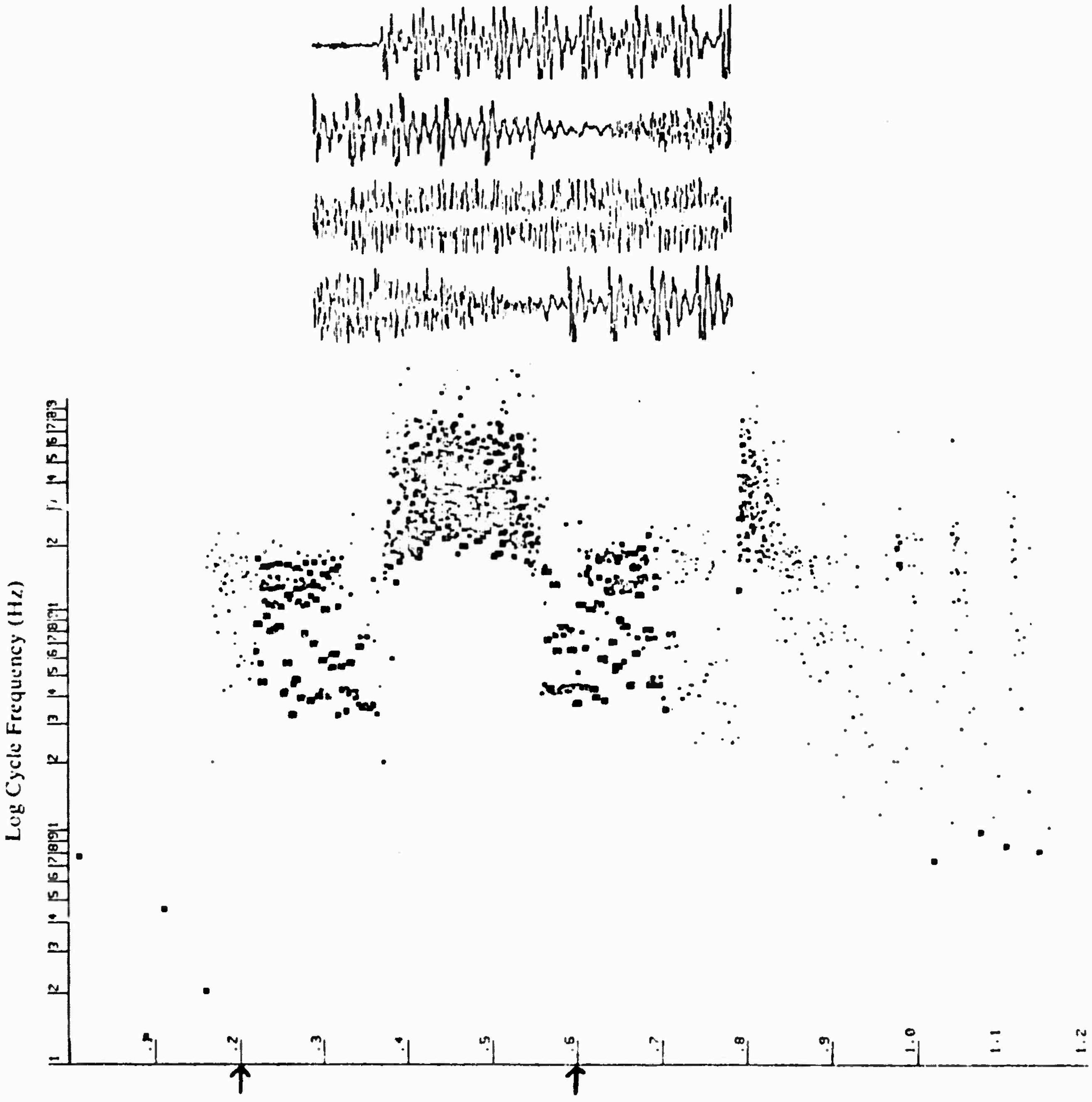


Figure 9: her shirt (σ,EH)

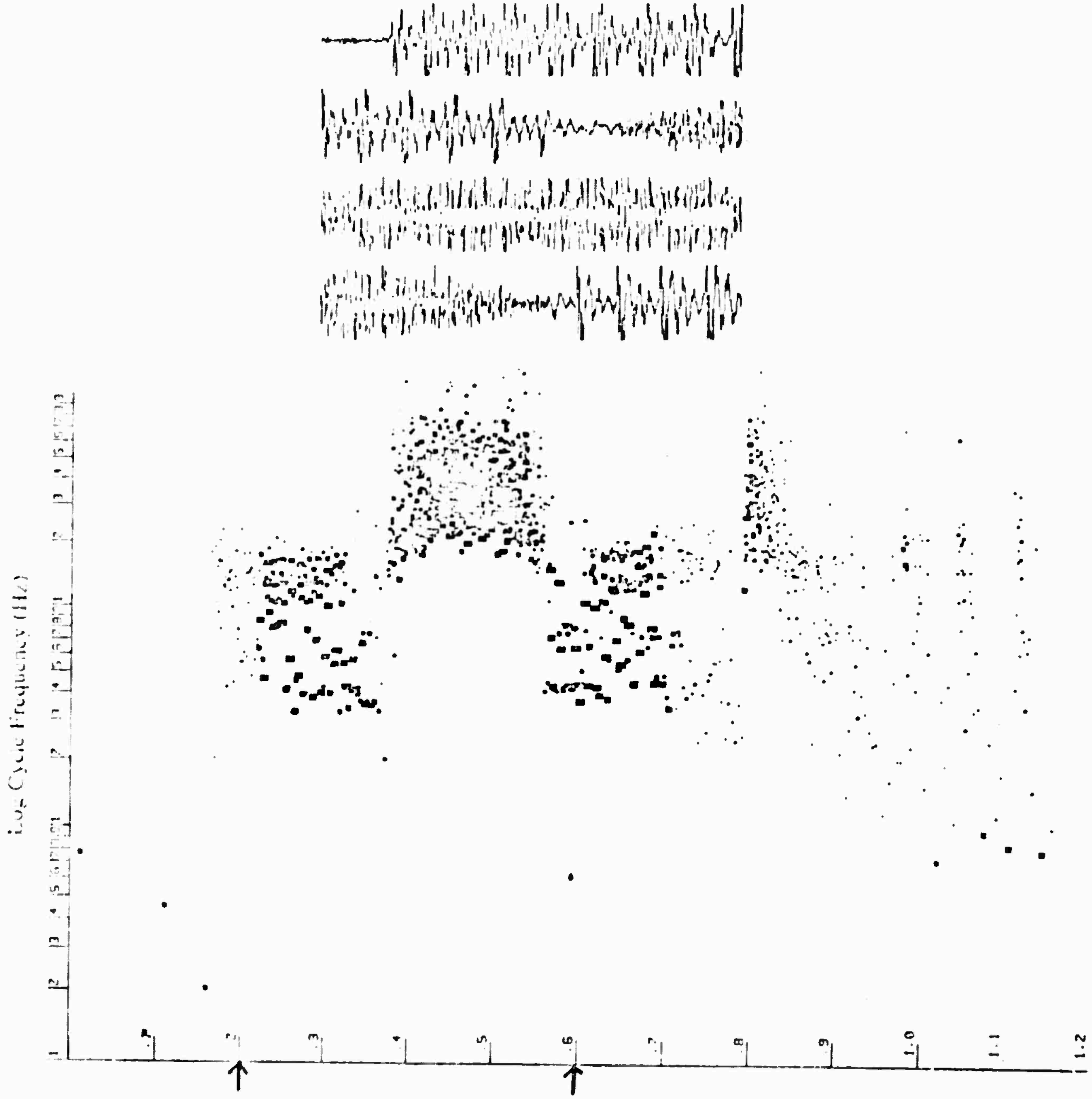


Figure 9: her shirt (σ,EH)

STOP CONSONANTS

The stop consonants are the set /p/, /t/, /k/, /b/, /d/, and /g/. Acoustically, stop consonants typically have a pause portion followed by a higher frequency region which represents the stop consonant release portion, plus aspiration, if present. A voiced stop consonant has a low frequency voicing region just preceding the pause portion. Often these lower voicing frequencies are sustained throughout part or all of the release-aspiration region as well. It is not uncommon for the pause segment to be completely omitted in a voiced stop consonant.

As the waveform transitions from prior context, or the initial voicing region in voiced stop consonants, into the pause portion, the cycle-frequency, amplitude, and microstructure measures sharply decline. This pause portion consists of only one or a few cycles, and the cycle-frequencies are quite low, usually less than 100 Hz. This abrupt drop in cycle-frequency is visually quite apparent in the LIP plots. Since the pause cycles are often of very low amplitude, and therefore are represented by very small dots on the LIP plots, we have visually enhanced them by automatic replacement of them with an asterisk symbol, "*". Next, as the waveform transitions abruptly into the release-aspiration region, both cycle-frequency and microstructure increase sharply as does amplitude, which nonetheless at its peak value generally remains well below the amplitude value for stressed vowels and most unstressed vowels. Where aspiration is clearly present, the transition from release to aspiration is usually quite smooth, though often with the cycle frequency and amplitude values gradually decreasing.

In the LIP plots shown here, pause cycles are marked "p", the release-aspiration region by "r", and the initial voiced region of voiced stop consonants by "v". The following example is of the /t/ (Fig. 10) in the utterance "the till" (EHP). Here the pause cycle has a cycle-frequency of 82 Hz. The release-aspiration portion has an average cycle-frequency value exceeding 4000 Hz and a duration of 68.3 msec. which is longer than is usually found in connected speech.

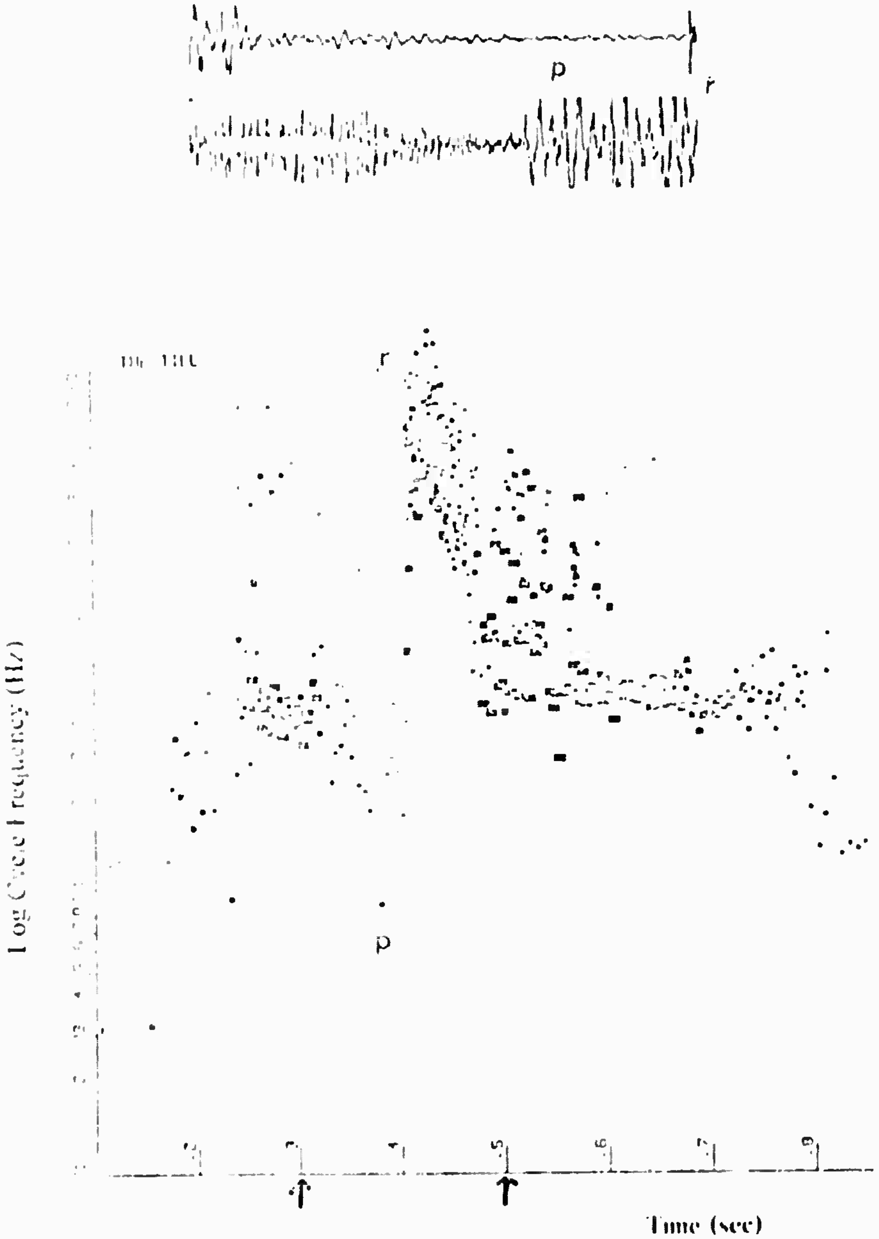


Figure 10: the till (σ,EH)

Time-domain analysis also reveals the existence of several more subtle acoustic phenomena. These phenomena are often both short in duration and low in amplitude. They occur at phone boundaries and last for only one or a few cycles in the acoustic waveform.

The first of these is analogous to the transitional cycles previously described for fricatives. At the end of the release-aspiration region of the stop consonant, there is often, though not always, one or a few cycles which have lower cycle-frequencies than any of the other cycles in either of the acoustic segments immediately preceding and following this acoustic event. These transitional segments are marked as "t" in the LIP plots which follow.

The second phenomenon shall be referred to as a "stop preview". In the case of a stop consonant which is preceded by a vowel (and sometimes by othe phone types as well), the very end of the vowel is sometimes characterized by one or two cycles which have much higher cycle-frequency values than any of the other cycles which comprise the vowel. These stop previews are usually very low in amplitude. Their duration is almost always less than 1 msec and commonly less than .5 msec. In the LIP plots, these are marked as "sp". The physical cause for these stop previews is not known. It is possible that they constitute a stop "closure" phenomenon resulting from high frequency turbulence as the articulators close in preparation for the stop consonant.

The third phenomenon concerns the one or two cycles immediately preceding the stop preview. These one or two cycles are usually of large amplitude, but have a lower cycle-frequency value than any of the cycles immediately preceding. Only at the very beginning of the vowel are there other cycles with cycle-frequencies as low or lower than the cycles immediately preceding the stop preview. These "stop preview transitional cycles" are sometimes omitted when the stop preview is present. They are marked "spt" in the LIP plots.

Illustrative examples of all these phenomena are provided in the utterance "to do" (Fig. 11), (HN,9). Here the cycle-frequency of the stop preview transition is 299 Hz and the two cycles comprising the stop preview have cycle-frequencies of 2538 Hz and 2695 Hz, respectively. The duration of the stop preview is .77 msec. Immediately following is the voicing region with an average cycle-frequency of less than 200 Hz and a duration of 70.3 msec. The release-aspiration portion of the /d/ has an average cycle-frequency exceeding 1500 Hz and a duration of only 16.8 msec. Two transition cycles mark the boundary between the /d/ release and the following vowel. Their cycle-frequencies are respectively, 325 Hz and 351 Hz. All of these abrupt changes in cycle-frequency, amplitude, and microstructure, resolve the stop consonants into several unambiguously distinct acoustic regions.

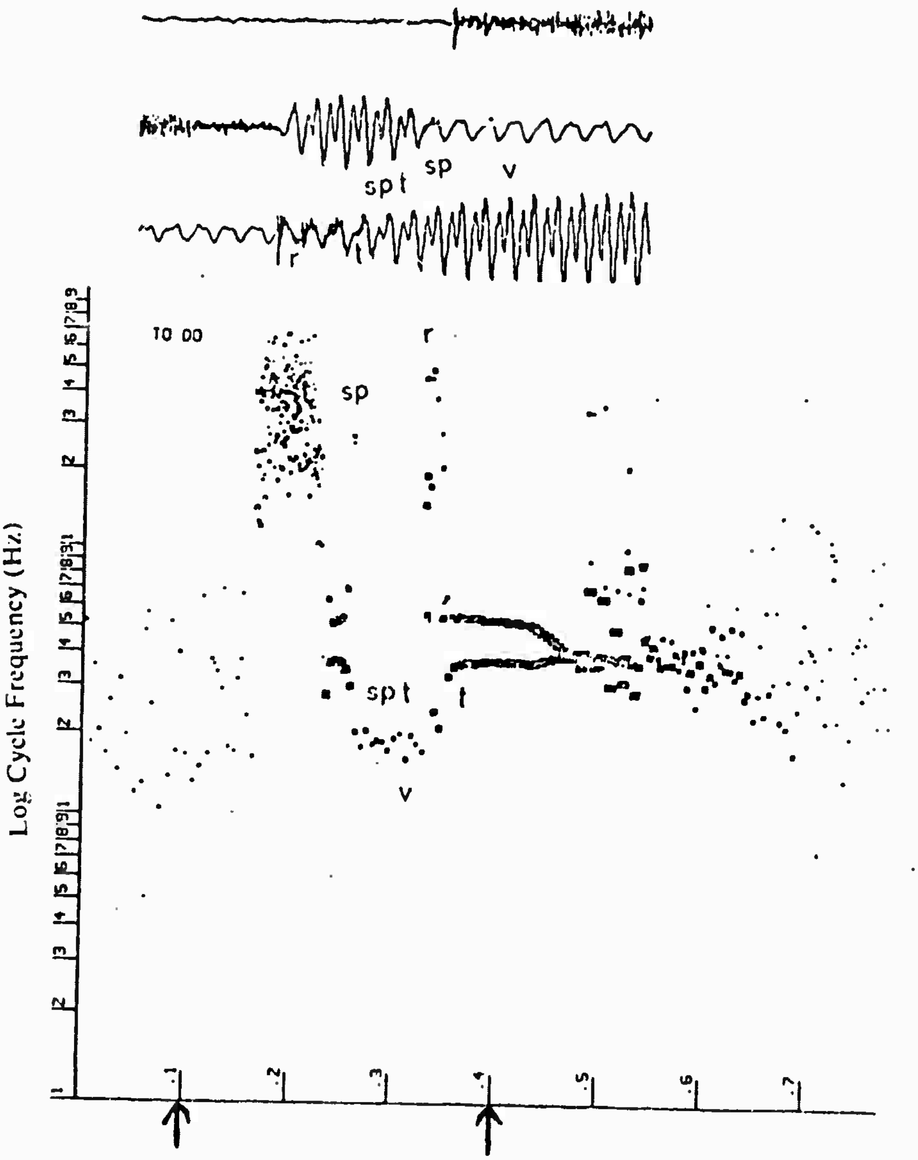


Figure 11: to do (♀,HN)

The following LIP plots provide representations of each of the other stop consonants. In these the different acoustics are individually marked but are analogous to those previously designated. These LIP plots show the /p/ (Fig. 12) in "display the" (LM,.), the /b/ (Fig. 13) in "the back lefthand" (RC,'), the /k/ (Fig. 14) in "has whitlockite" (BB,), and the /g/ (Fig. 15) in "he grows" (HN,').

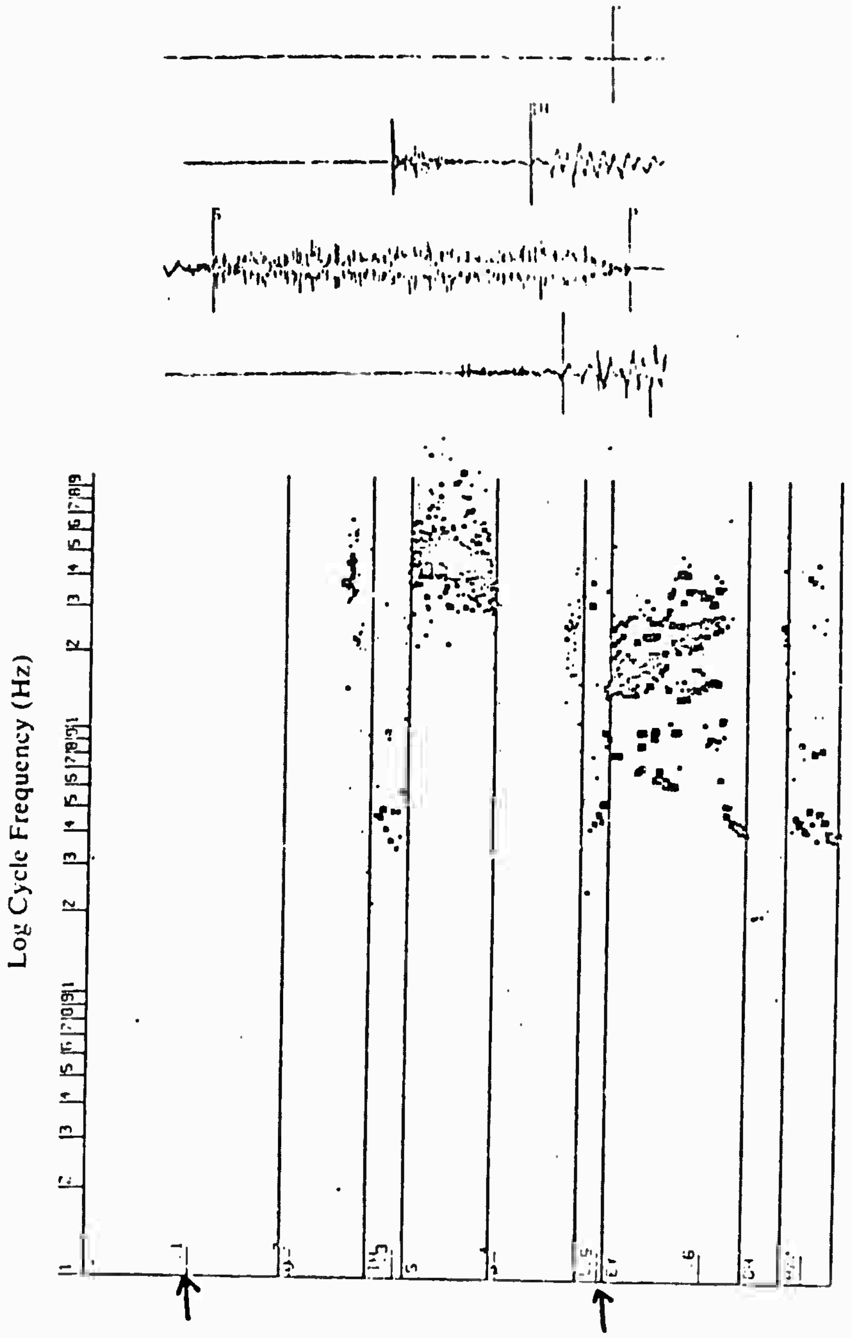


Figure 12: Display the (?,LM)

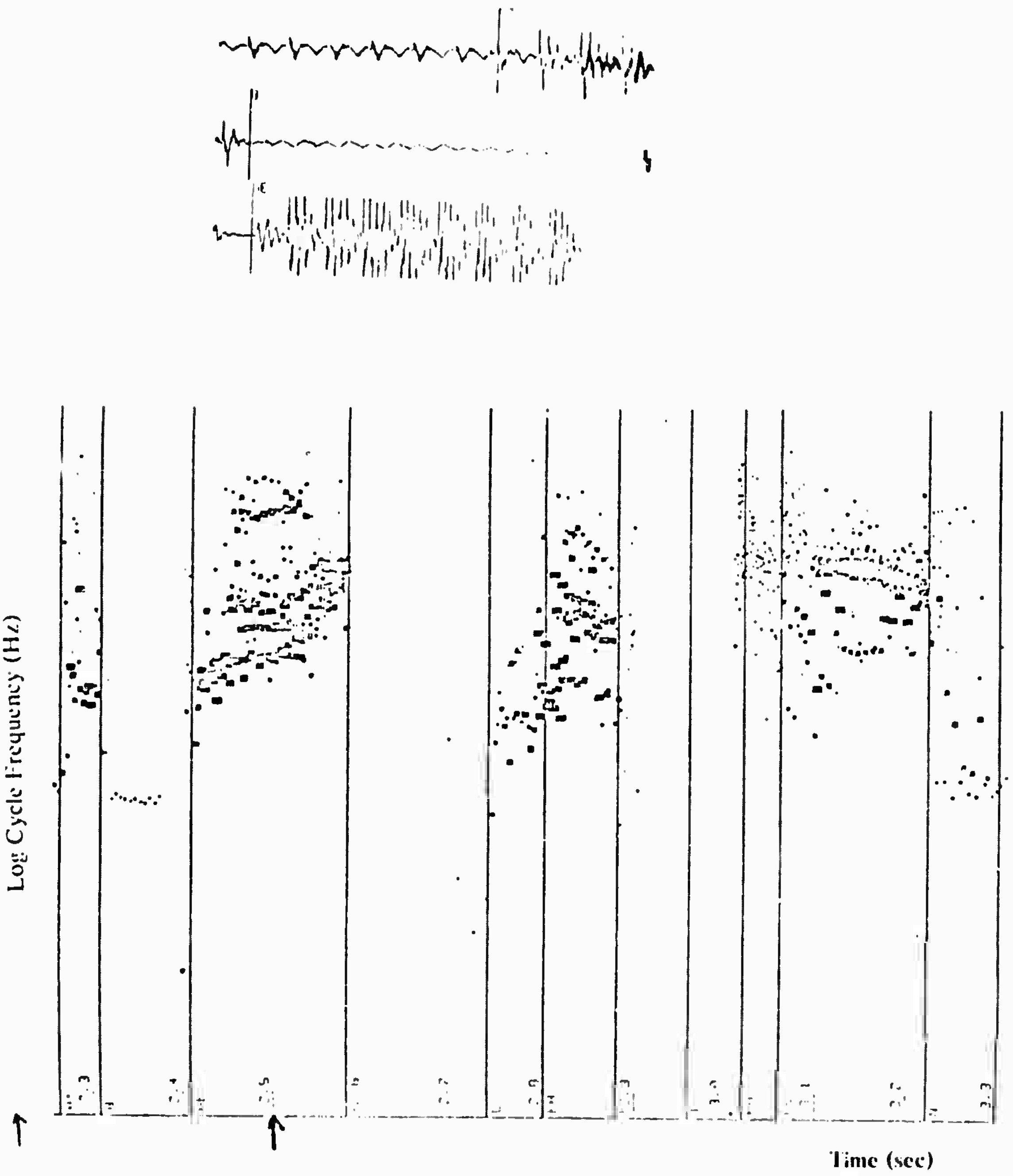


Figure 13: the back lefthand (c,[illegible]C)

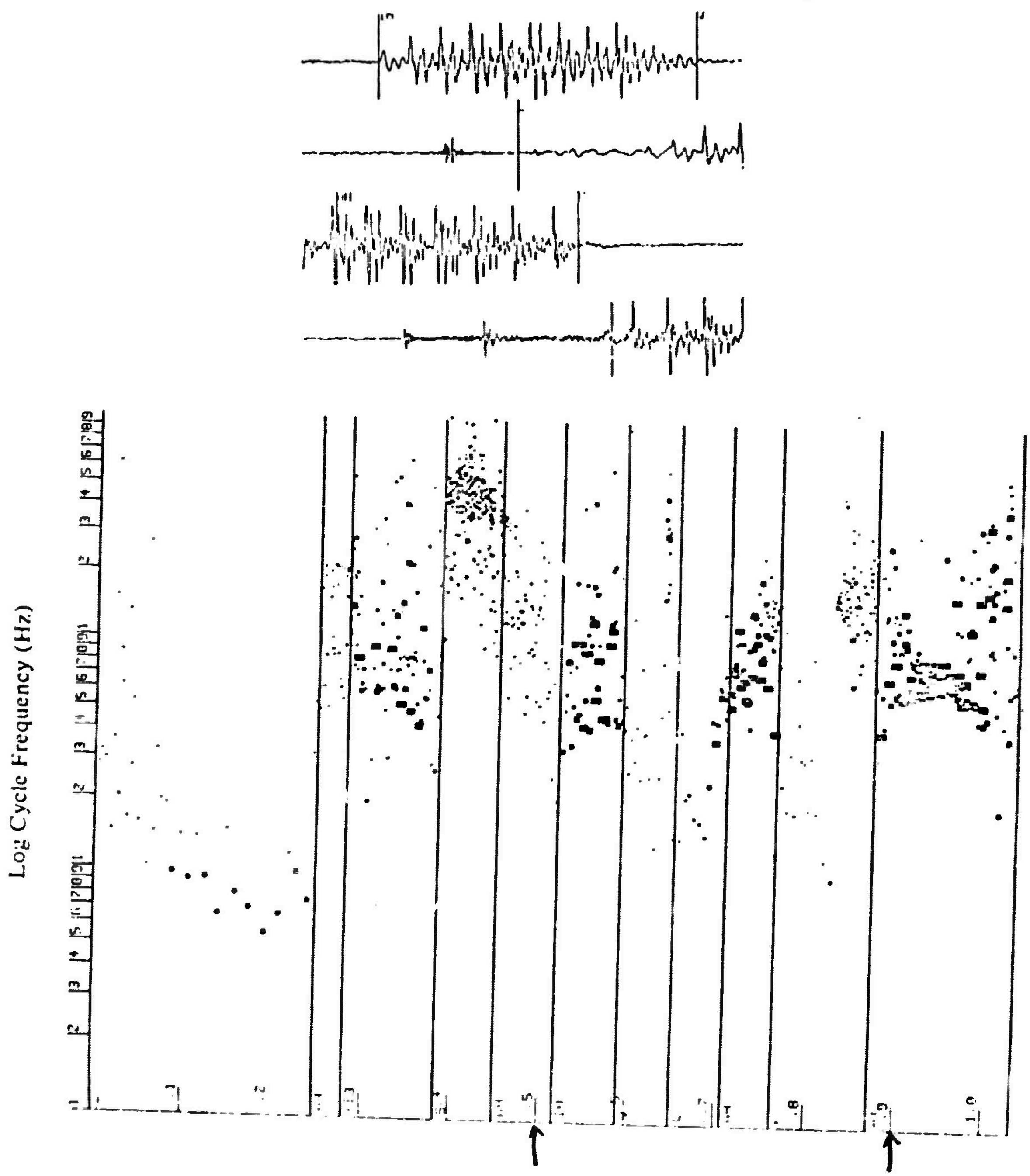


Figure 14: Has whitlockite (σ,BB)

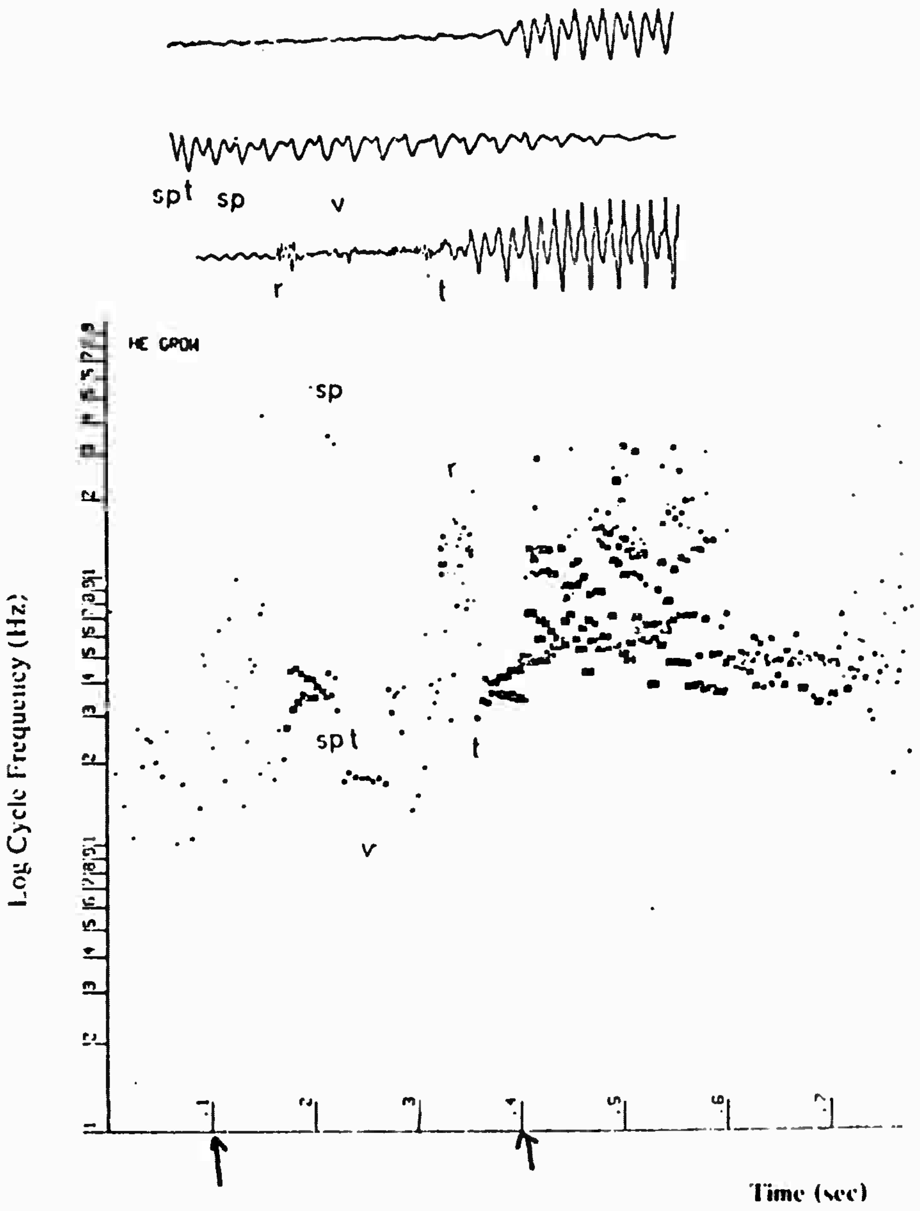


Figure 15: he grows (♀,HN)

ACOUSTIC-PHONOLOGICAL PHENOMENA

There are a variety of acoustic phonological phenomena which are commonly observed with time-domain analysis. Generally these phenomena are readily apparent in both the waveform and LIP plot. However, especially when such acoustic events are either very brief in duration or low in amplitude, or both, their existence is often much more visually evident in the LIP plots.

One very common phenomenon is the case where a fricative is characterized by a central region where the cycle-frequencies are lowered in relation to that phone's characteristic frequency. This central region may result from articulatory changes which occur in preparation for articulation of the following phone. In the following example, the phone of interest is a rounded /f/ (Fig. 16) in the utterance "no foe" (HN,9). Here the average cycle-frequency value for the initial fricated region is 1079 Hz, for the central region is 693 Hz, and for the final fricated region is 1148 Hz. In addition, the first fricated region is much greater in amplitude than the central and final fricated regions, which are about equal in amplitude. Two more examples follow with the /s/ (Fig. 17) in "pyroxine" (BB,) and the second /f/ (Fig. 18) in "from left to" (RB,).

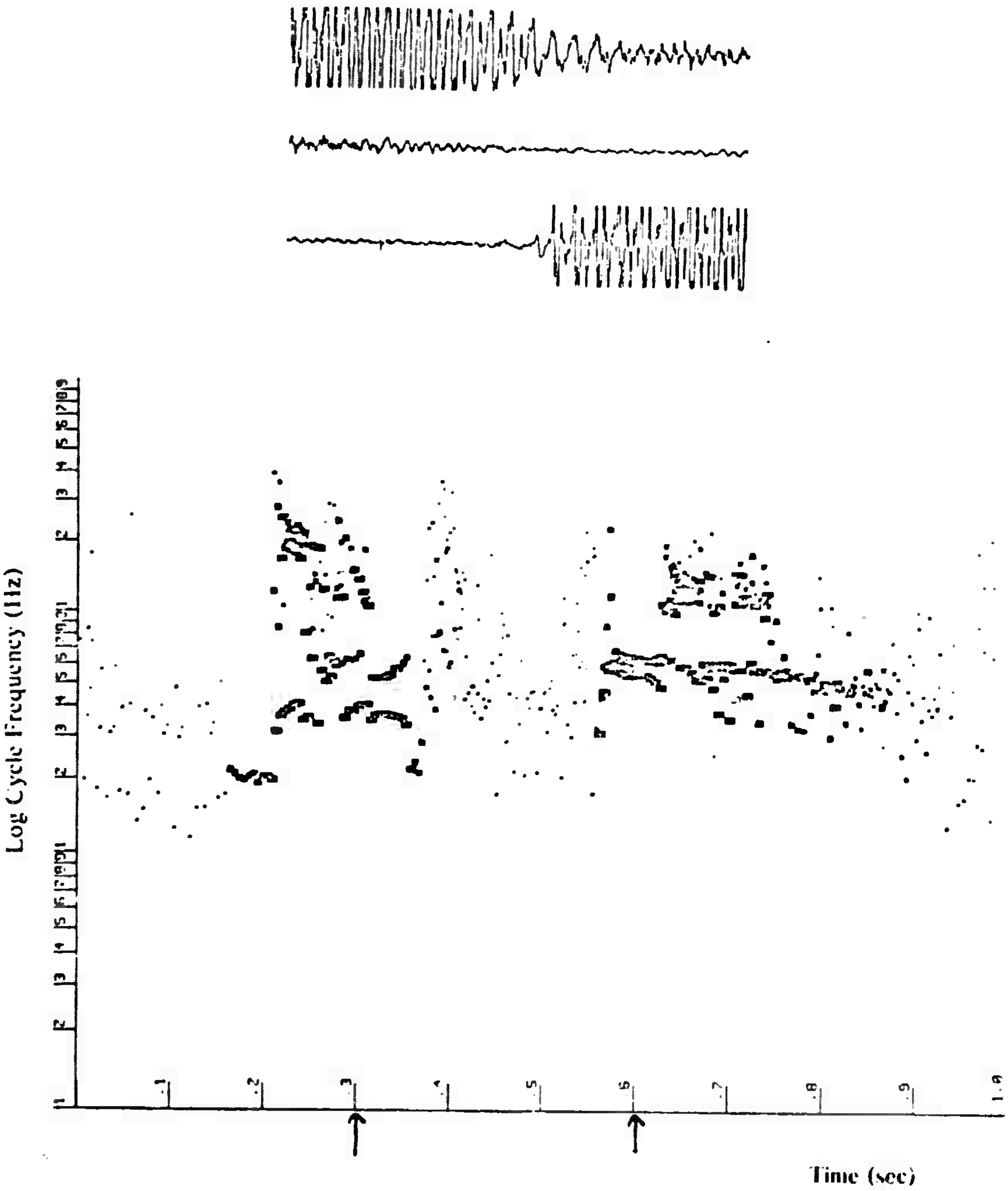


Figure 16: no foe (♀,HN)

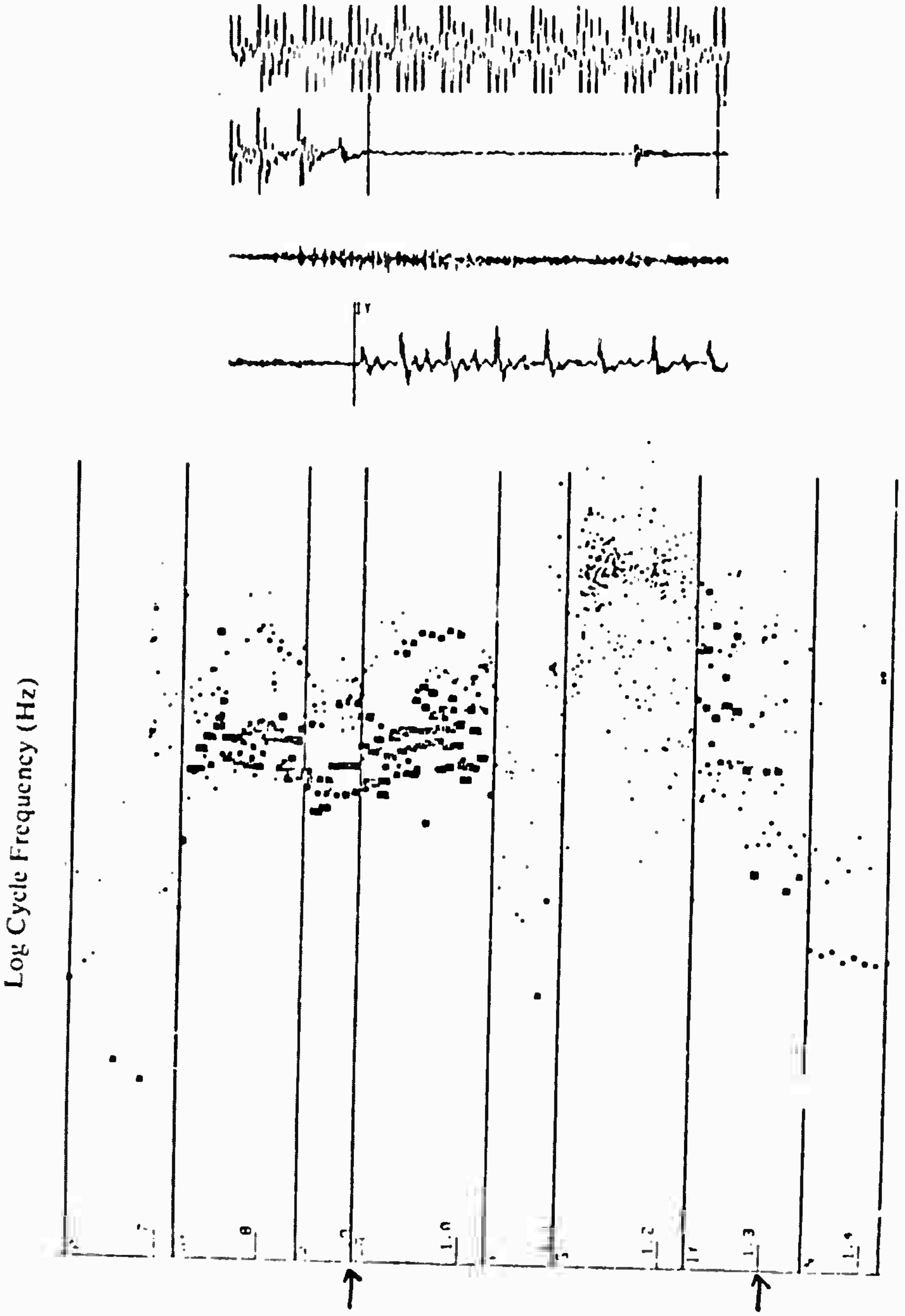


Figure 17: pyroxine (σ,BB)

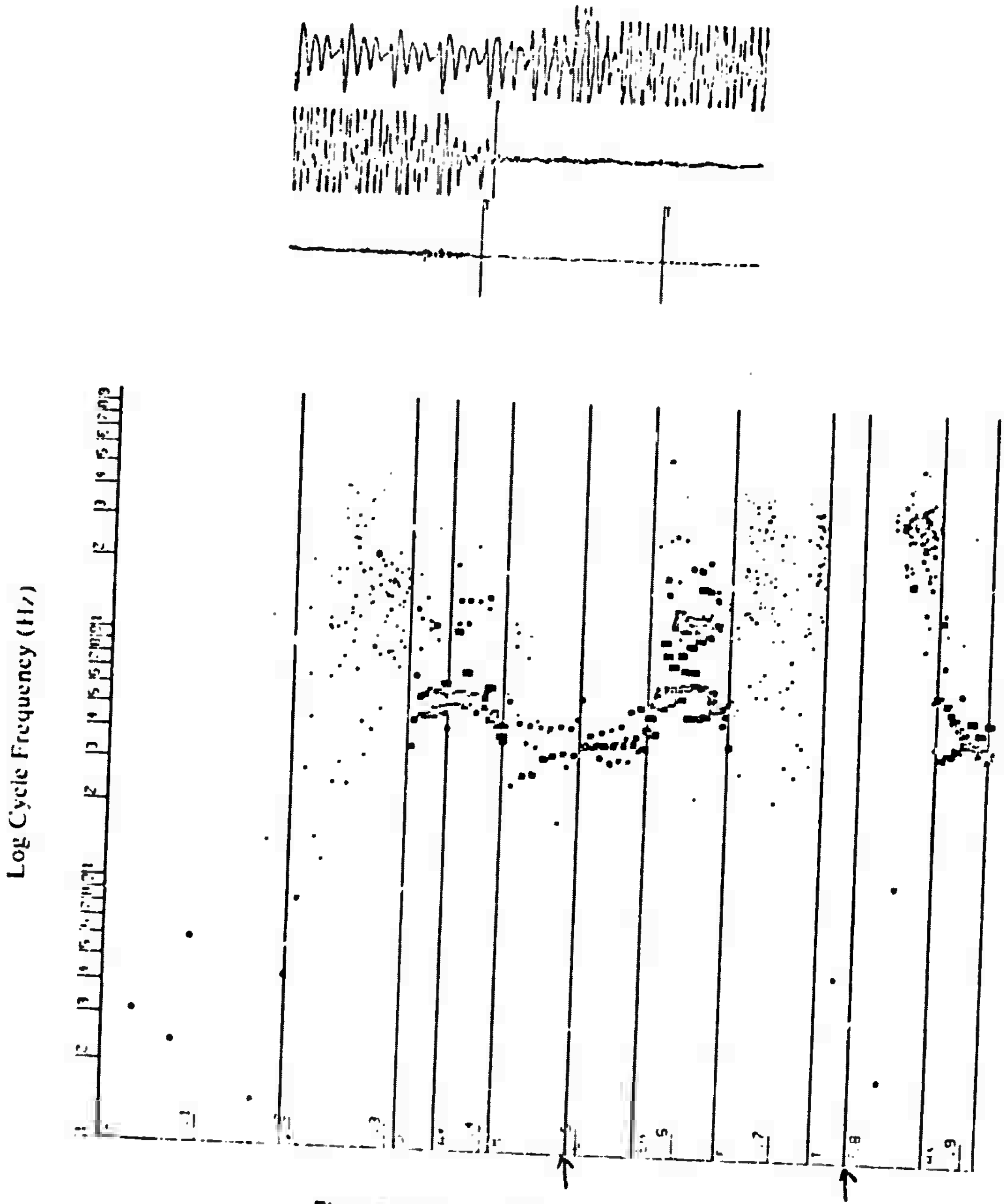


Figure 18: From left to (σ,BB)

The next phenomenon concerns the issue of the acoustic correlates of what are commonly referred to as "unreleased" stop consonants. Time-domain analysis reveals that many of the stop consonants which are phonetically transcribed by linguists as "unreleased" or "deleted", are better described as "minimally released". These are acoustically characterized by the usual pause cycle(s), followed by a very brief segment of high cycle-frequency energy which is analogous to a normal release-aspiration segment except for its short duration, and which is sometimes followed by the transition cycle(s) leading into the next phone. This very brief release-aspiration segment consists of only a few cycles, often just one or two cycles, where the entire duration of this portion ranges from less than 1 msec to more than 6 msec. The temporal sequence of acoustic events characterizing these minimally released stop consonants is essentially identical with that for normally released stop consonants, except for durational aspects. The few cycles with high cycle-frequencies remaining in minimally released stop consonants are generally insufficient for reliable identification of the stop consonant. However, the information that a stop consonant has occurred, and whether or not it was voiced, does remain in most cases. The following example shows such a minimally released stop consonant, the /d/ (Fig. 19) in "would give" (HN,7). Here the release-aspiration segment is comprised of only 2 cycles, with a total duration of 1.4 msec, and is preceded by a normal voicing region.

The following LIP plots provide several more examples of this same phenomenon, very commonly found in connected speech, especially with the first of two successive stop consonants. Acoustic observations of a very brief stop consonant often indicate that another stop consonant will immediately follow. These examples show the /b/ (Fig. 20) in "tub took" (E,'), the /k/ (Fig. 21) in "spectrogram" (LM,), and the /p/ (Fig. 22) in "stoop to" (EH,).

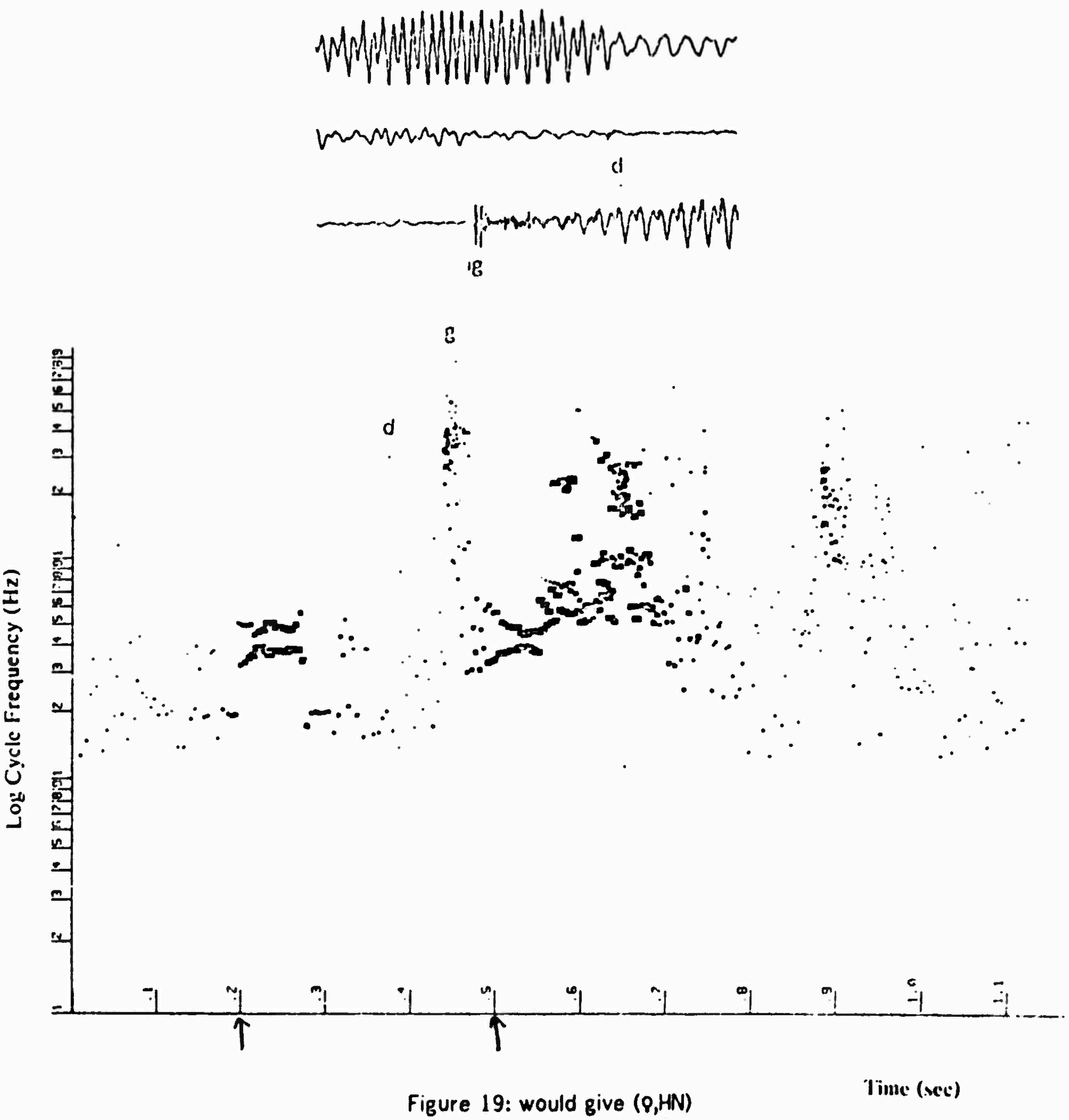


Figure 19: would give (ϙ,HN)

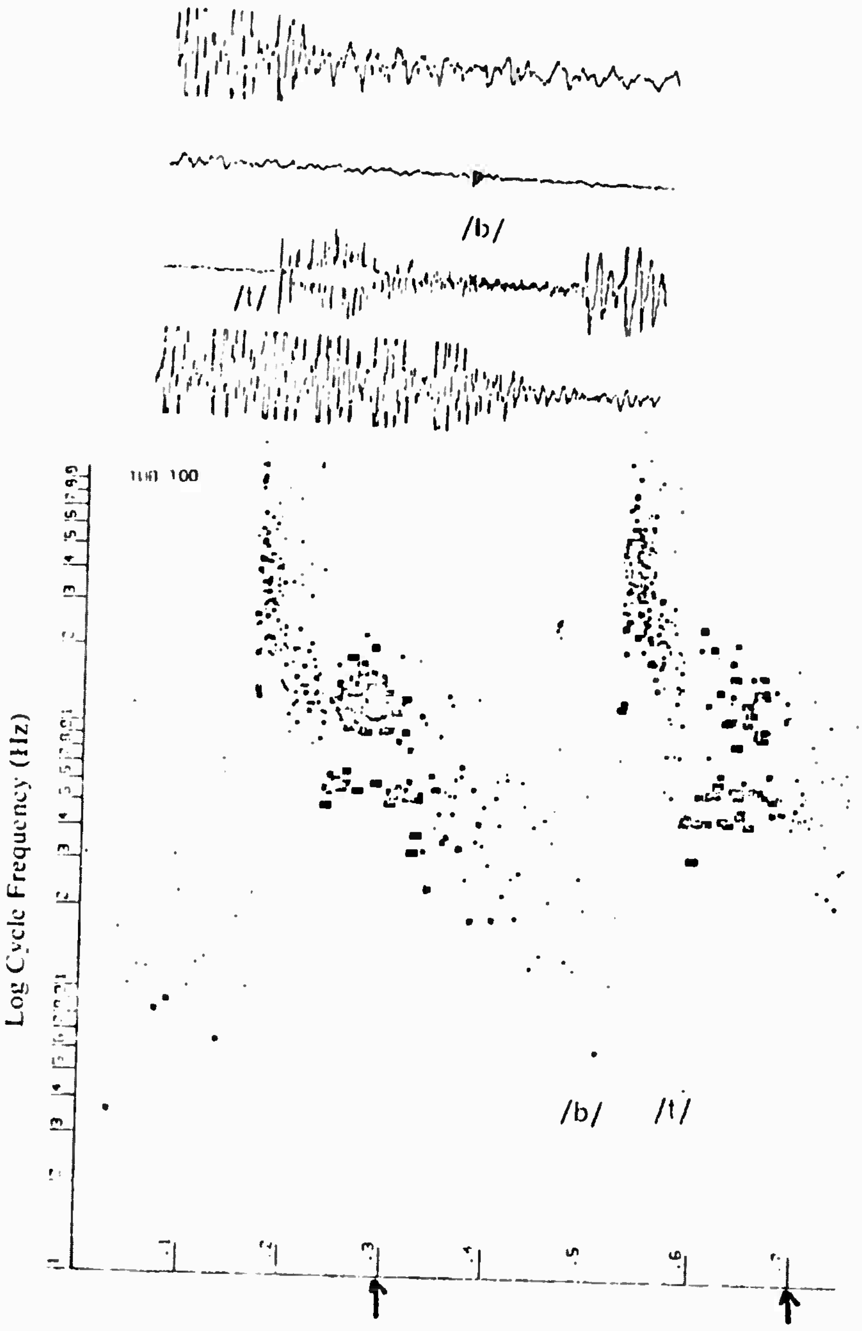
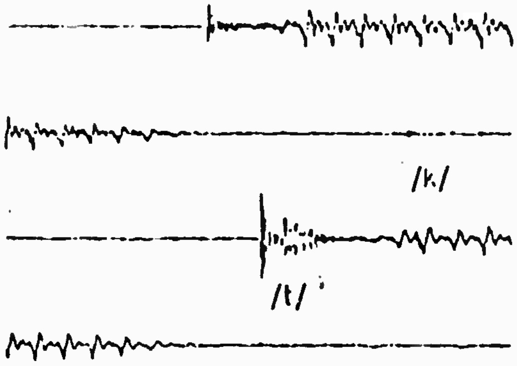


Figure 20: tub took (σ,EH)



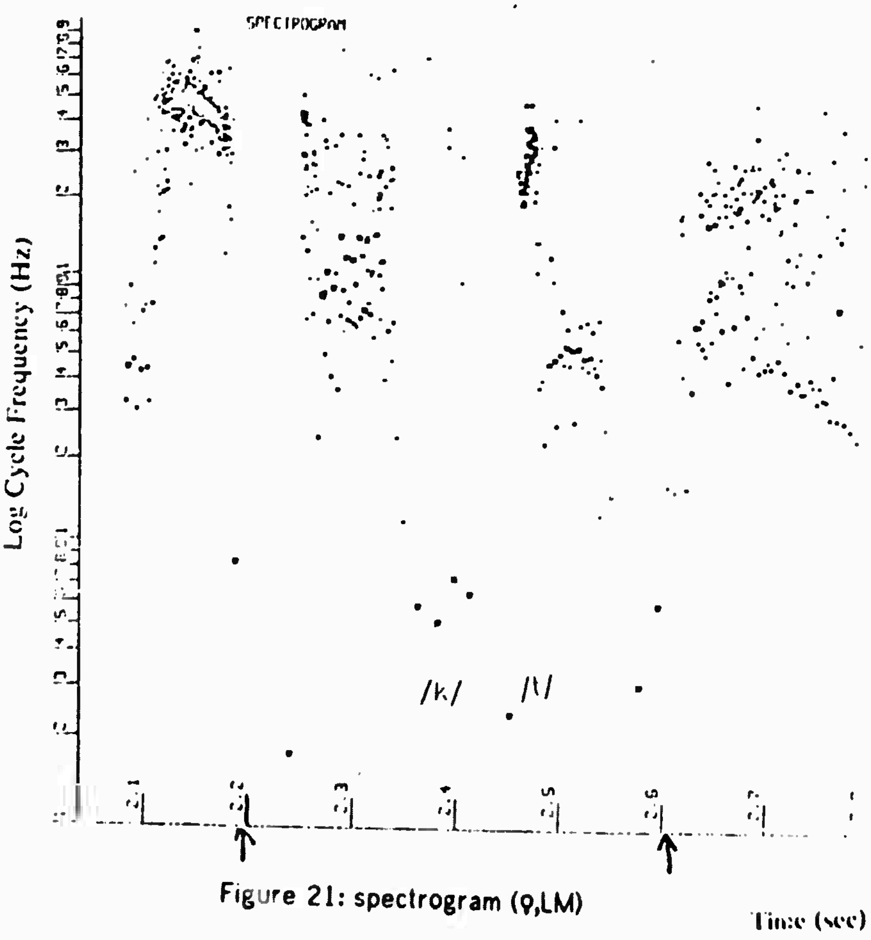


Figure 21: spectrogram (9,LM)

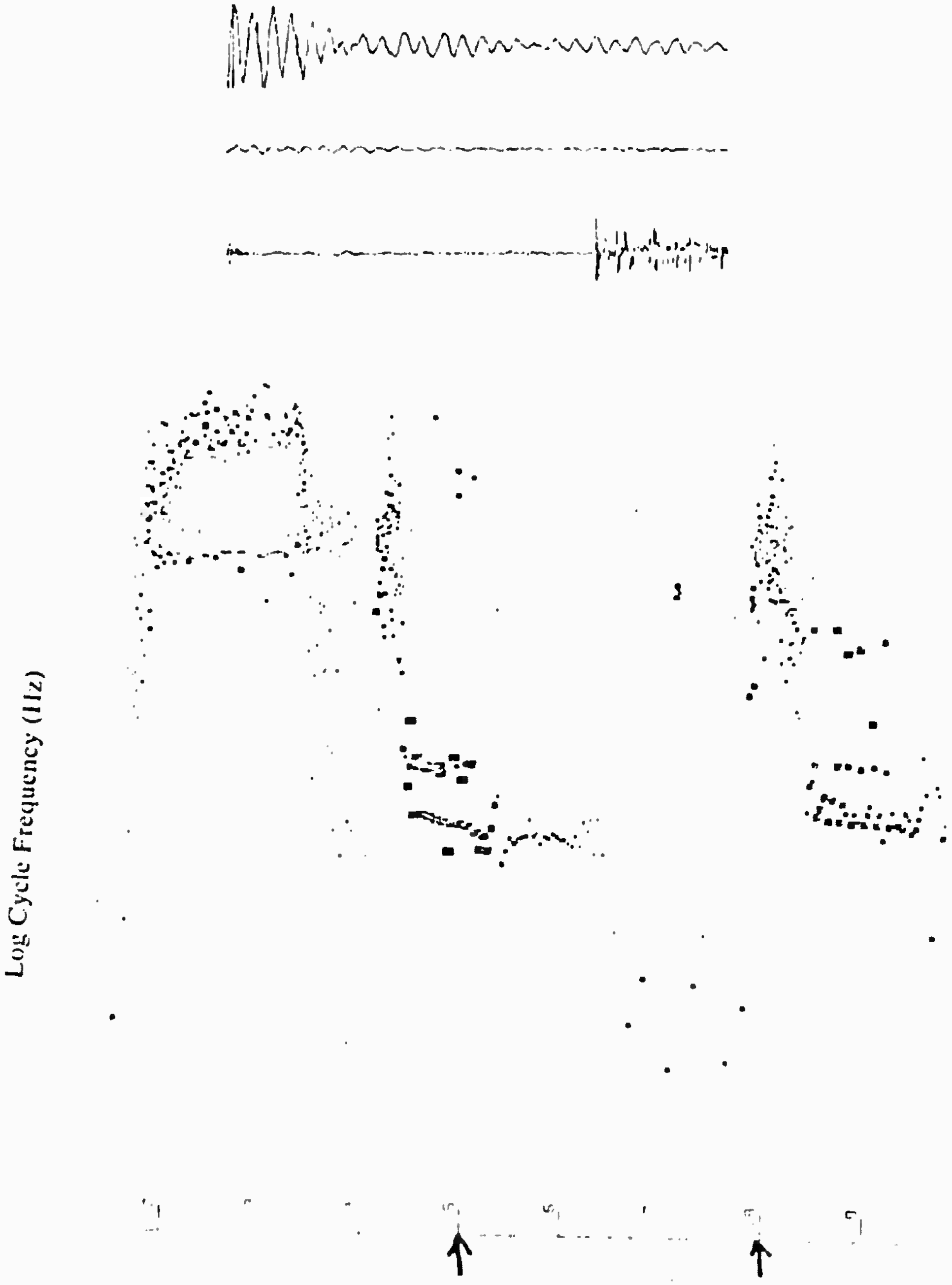


Figure 22: stoop to (σ,EH)

Another phonological phenomenon relates to the occasional insertion of an extra stop consonant. This occurs when a syllable ends with a stop consonant and the next syllable begins with a vowel (or sometimes even a liquid or nasal), even when a substantial interword pause boundary separates the two syllables. The speaker will often articulate a normally released stop consonant at the end of the first syllable as expected, but then repeats this same stop consonant in a briefer less intense form, when he begins the next syllable. We refer to this consonant doubling phenomenon as "gemination". When this happens, it is not perceptually obvious to a listener that a second stop consonant has been inserted by the speaker. Although this phenomenon occurs frequently, it is not so obvions from visual displays if the inter-syllable or inter-word pause is less than about 30 msec. In the following utterance (Fig. 23) "about Israel" (JB9), the /t/ in "about" is repeated, even after a long interword pause of .18 sec, just before the initial vowel in the word "Israel". In this example, both acoustic manifestations of /t/ are very similar. As compared with the first /t/, the second or inserted /t/ has a duration of 38% and an average amplitude of 82%. As can be seen from the LIP plot, the component cycle-frequencies of both are very similar, although slightly higher for the second /t/, which is followed by a high front vowel. Generally, where the preceding and following contexts for such a stop consonant are very different, the acoustic manifestations of that consonant may also differ substantially due to differing effects of coarticulation. Specific coarticulation effects on the acoustic characteristics of stop consonants will be discussed in Chapter IV.

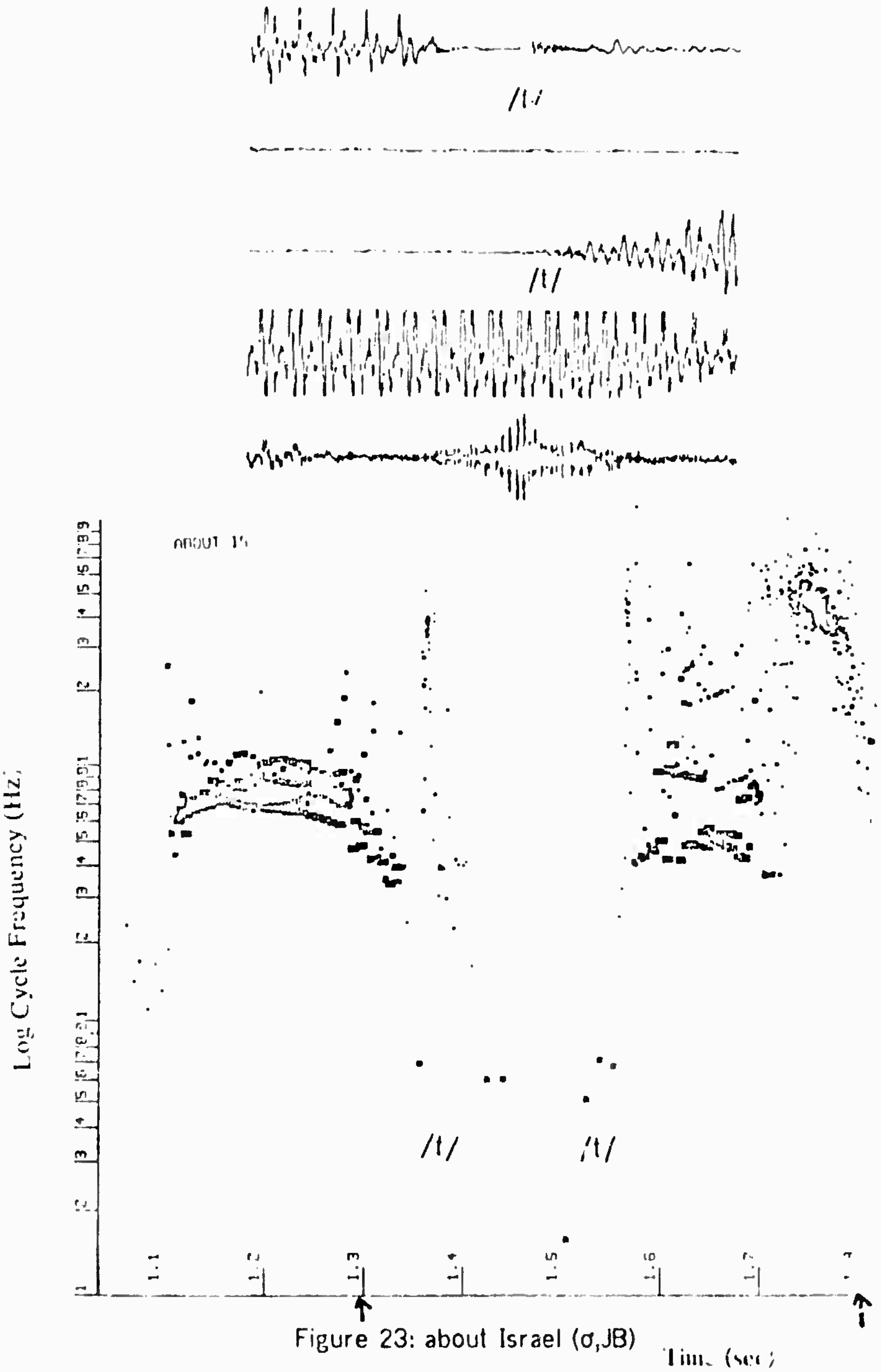


Figure 23: about Israel (σ,JB)

This phenomenon of the inserted stop consonants should not be confused with a normal glottal stop which commonly precedes an utterance initial vowel and differs substantially from a stop consonant. In contrast to the stop consonants, the release of a normal glottal stop is not composed of a well-defined and sustained high cycle-frequency region, but is characterized by a diffuse region of mixed cycle-frequency components. Shoup [S3] has suggested, however, that this acoustic occurrence of a second stop consonant may be produced by releasing a glottal closure while the articulators remain at or near the same position assumed during the preceding /t/ release. In order to determine the physical position of the articulators for this and other speech phenomena, major studies must be performed which observe the articulator positions and movements with a precise synchronization of the speech waveform.

Several more examples are shown in the following LIP plots in the utterances "berthed all" (Fig. 24,EH,G), "lug more" (Fig. 25,HN,[illegible]), and "phonemic Labels" (Fig. 26,LM,[illegible]), where /d/, /g/, and /k/, respectively, are acoustically each manifested twice.

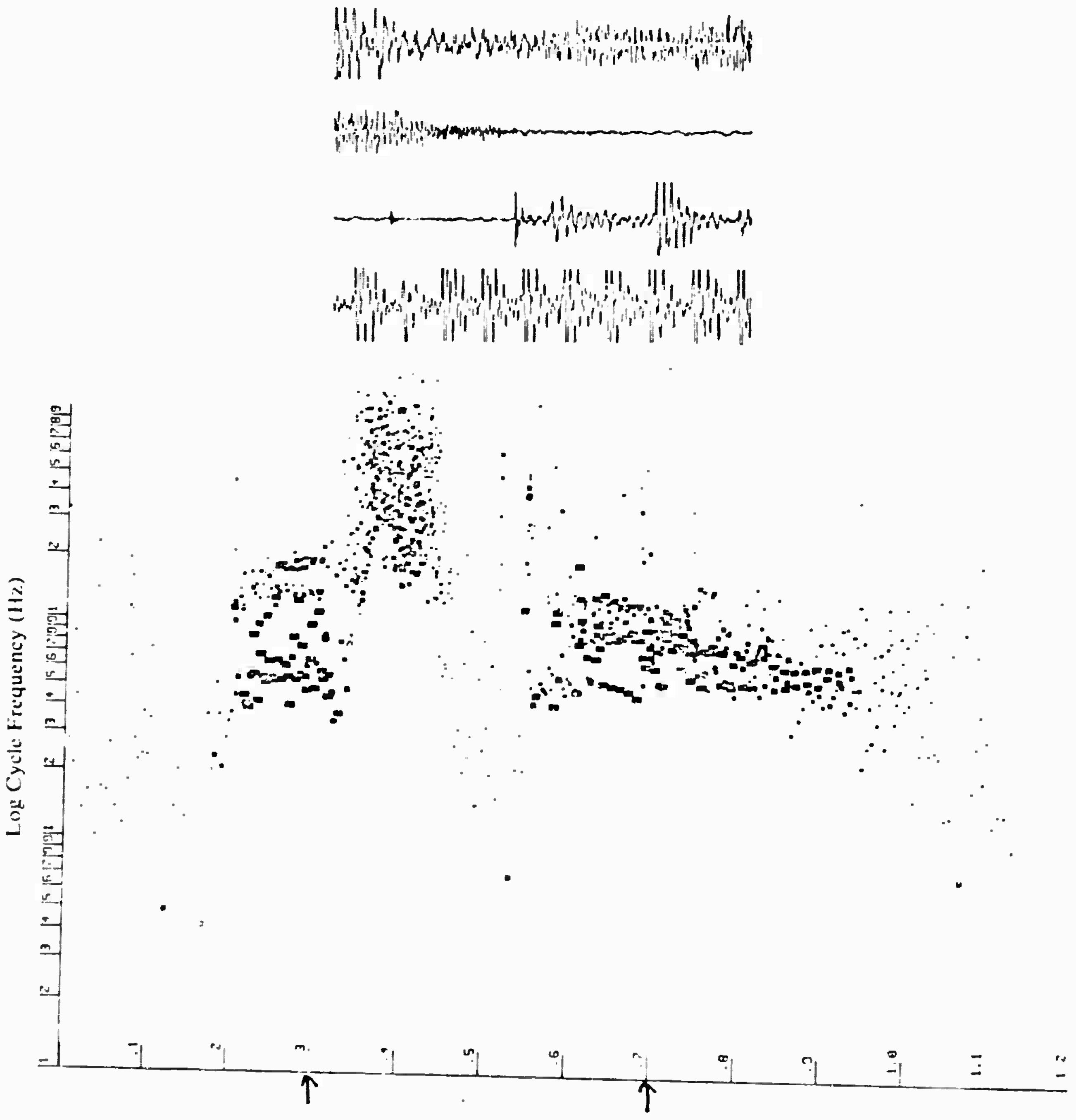


Figure 24: berthed al' (σ,EH)

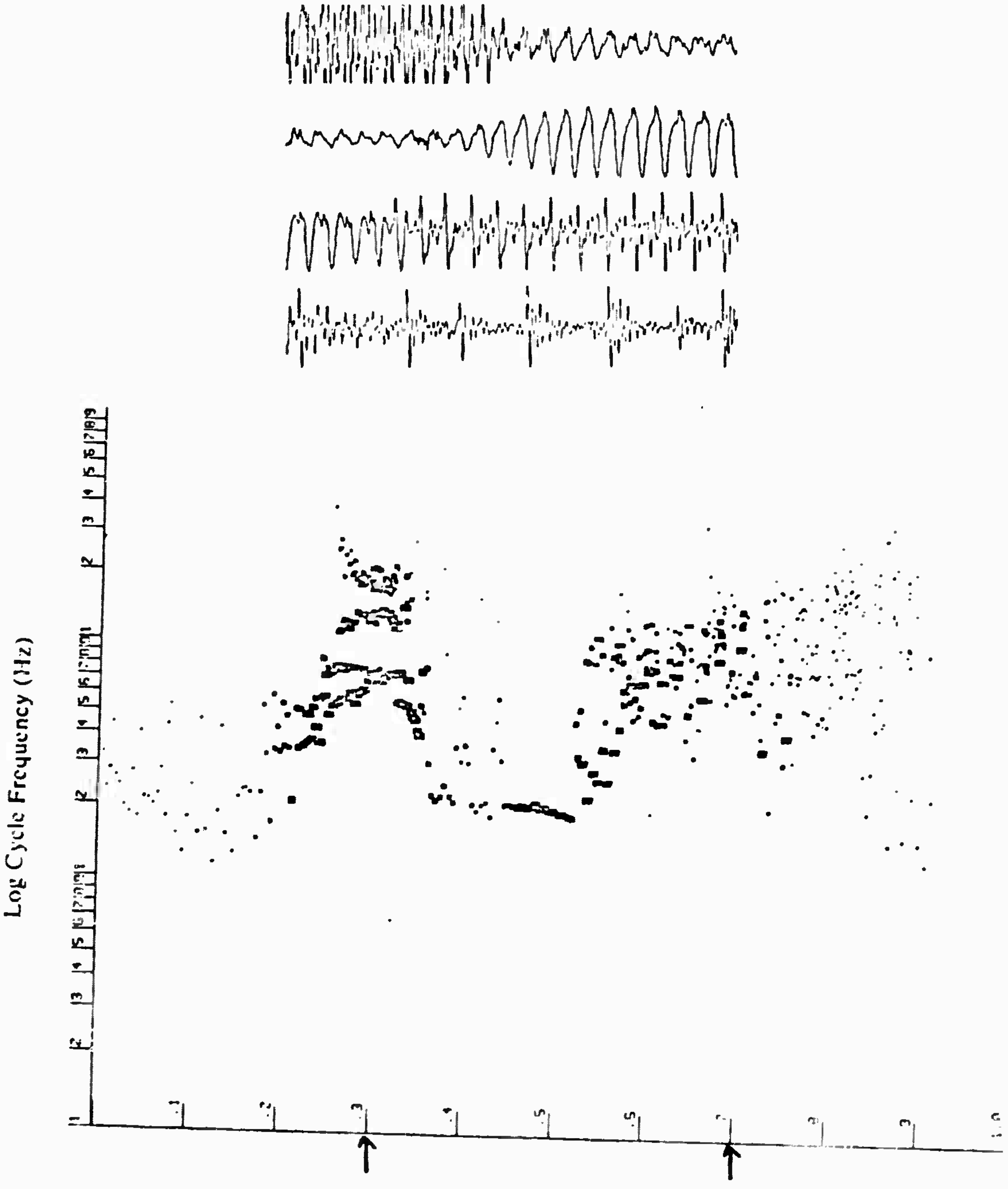
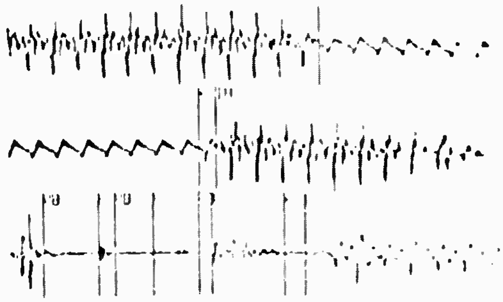


Figure 25: lug more (♀,HN)



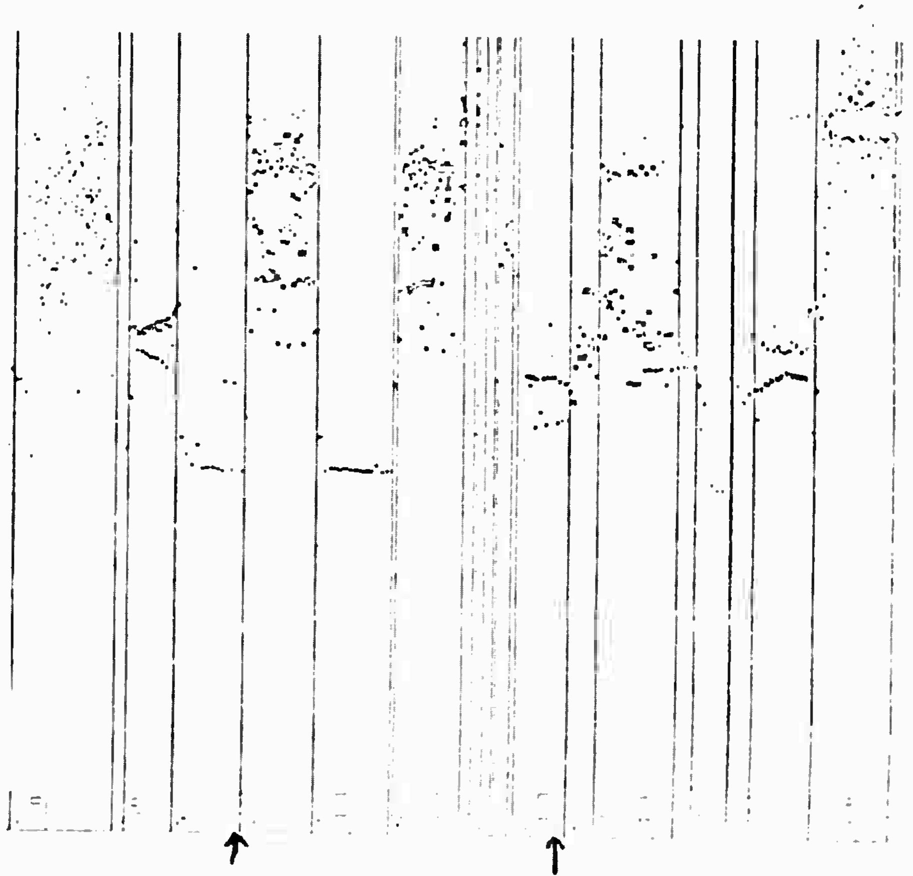


Figure 26: phonemic labels (ϙ,LM)

The Other Phones

Although our investigations have centered on the acoustic characteristics of fricatives and stop consonants, we note here general observations on some of the other phone classes. The characteristics cited occur commonly throughout our data sets.

VOWELS

The vowels, especially the stressed vowels, are very large in amplitude. Generally cycle-frequency and microstructure measures are somewhat variable throughout any given vowel although they tend to be higher for front vowels, as would be expected. The LIP plots of vowels sometimes appear as very regular patterns, but often appear highly irregular. The former case is demonstrated in the utterance "to sue" (Fig. 27,JA♂). Here the regular double-line pattern appears superficially to represent formant structures, although, of course, that is not the case with this time-domain technique. In these two vowels, each pitch period contains two major cycles of different frequencies, which are relatively consistent from one pitch period to the next, hence the double line.

Vowels often appear much more irregular in LIP plots. This especially occurs for high front and mid vowels where relatively large amplitude microstructure rides on the lower cycle-frequency components. This means that the individual periods of the major cycles within a pitch period are often increased or decreased by one or more of the higher cycle-frequency components riding on them. This kind of irregular pattern is exemplified in the LIP plot of "the gift" (Fig. 28,JA♀).

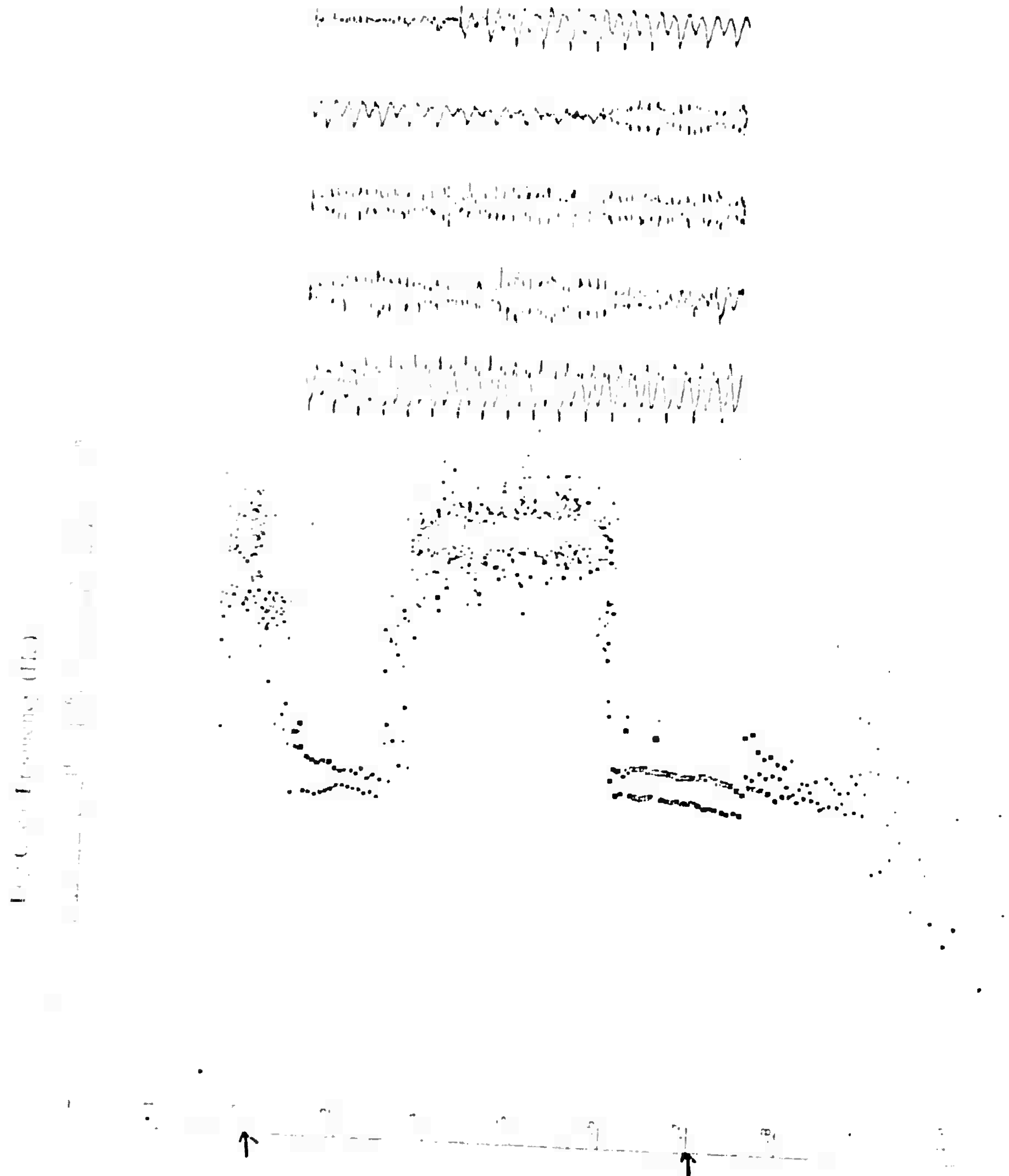


Fig. 27: to sue (σ,JA)

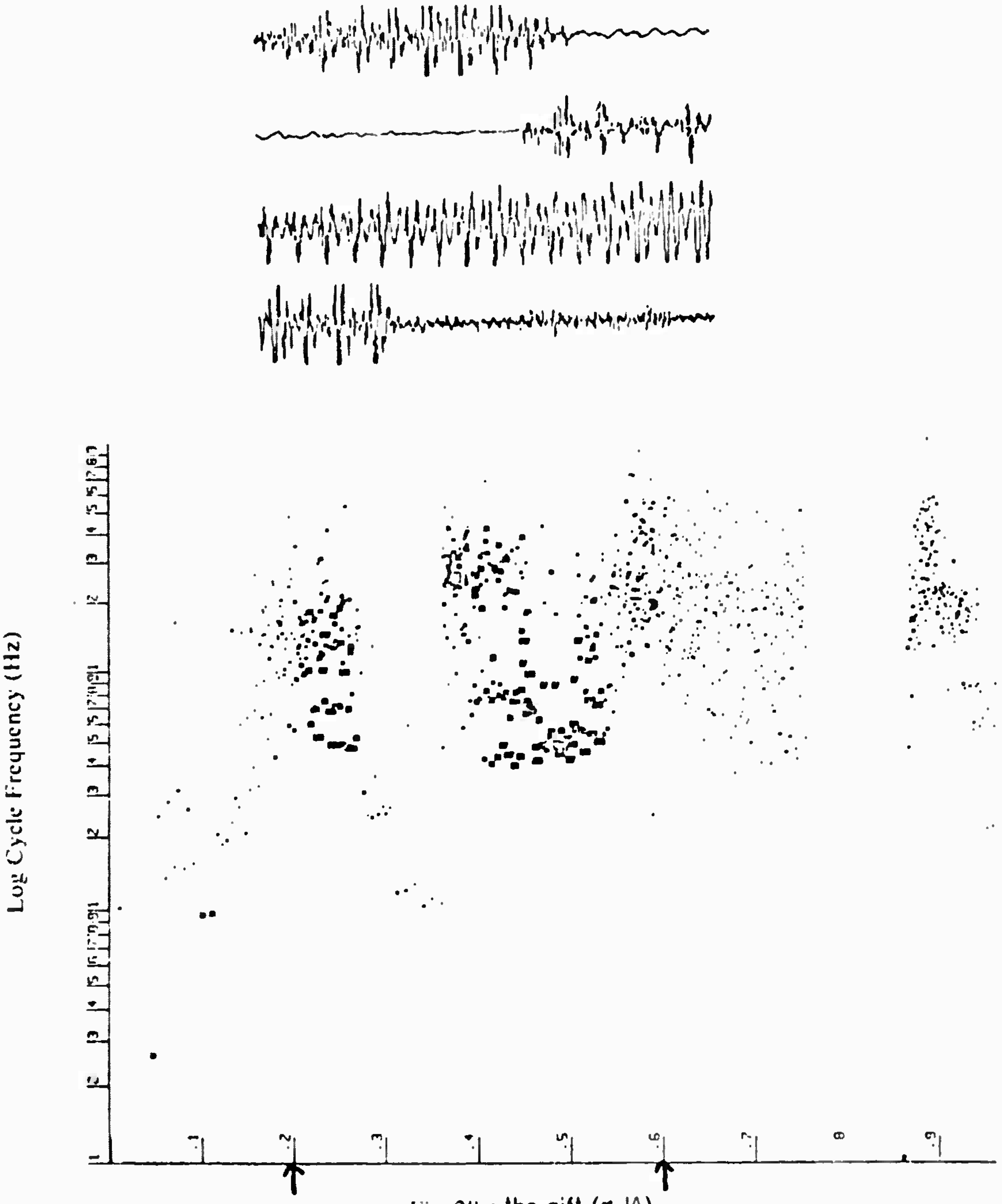


Fig. 28 : the gift (σ,JA)

Sometimes, especially for vowels of long duration, a transient event occurs during the last half or third of the vowel. This event consists of one or two unusually high cycle-frequency cycles which seem to occur just prior to an increase of irregularity in the vowel pattern as seen in the LIP plot. At this break in the vowel pattern, amplitude often decreases quite rapidly. General degradation of the vowel may be occurring during this end portion. If this is the case, then vowel identification would presumably be more reliable when based on acoustic observations of the vowel region prior to such a break. This phenomenon is illustrated in the vowel pattern in both vowels in the utterance "the key" (Fig. 29,HN,♀). The high cycle-frequency dots occurring at this vowel break, are marked "b"

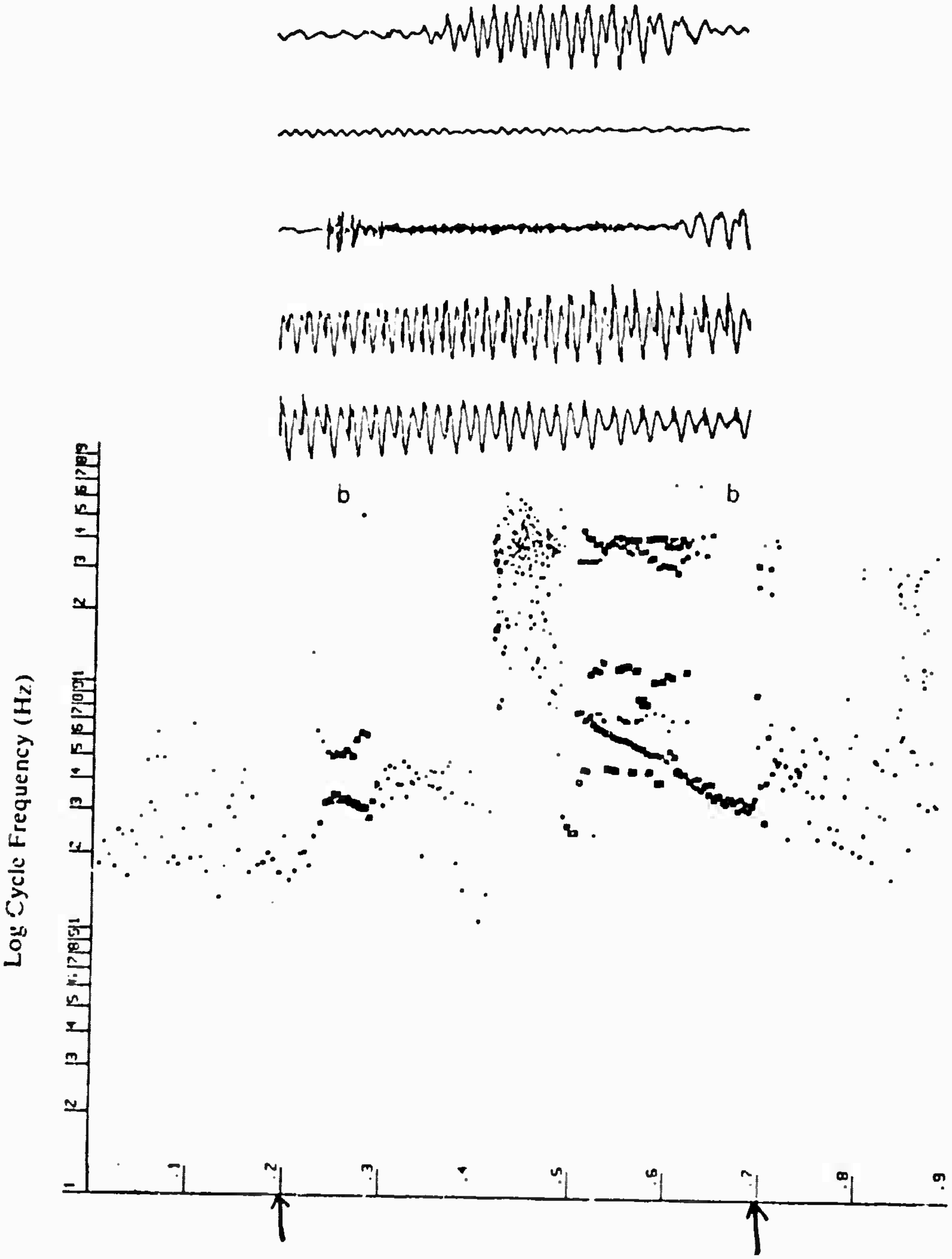


Fig. 29: the key (♀,HN)

LIQUIDS

The acoustic characteristics of liquids have not been thoroughly studied in the time-domain although we know that, taken as a class, liquids are quite variable with respect to cycle-frequency, amplitude, and microstructure. Of all the phone classes, liquids are hardest to segment out of the speech stream, by hand or automatically. In contrast to most other phones, liquids often start and end quite gradually with segue regions. There is, of course, no *a priori* reason for assuming that sharp boundaries should always exist. Two examples follow which illustrate some boundary ambiguities. First is /l/ (Fig. 30) in "week old" (JA,O). Next is the /r/ (Fig. 31) in the word "tridymite" (BB,C). A segmentation line has been inserted here where the juncture between the /t/ and /r/ was best approximated. In contrast, we also find less ambiguous junctures too. An illustrative example appears for the /r/ (Fig. 32) in the utterance "week ran" (JA,C).

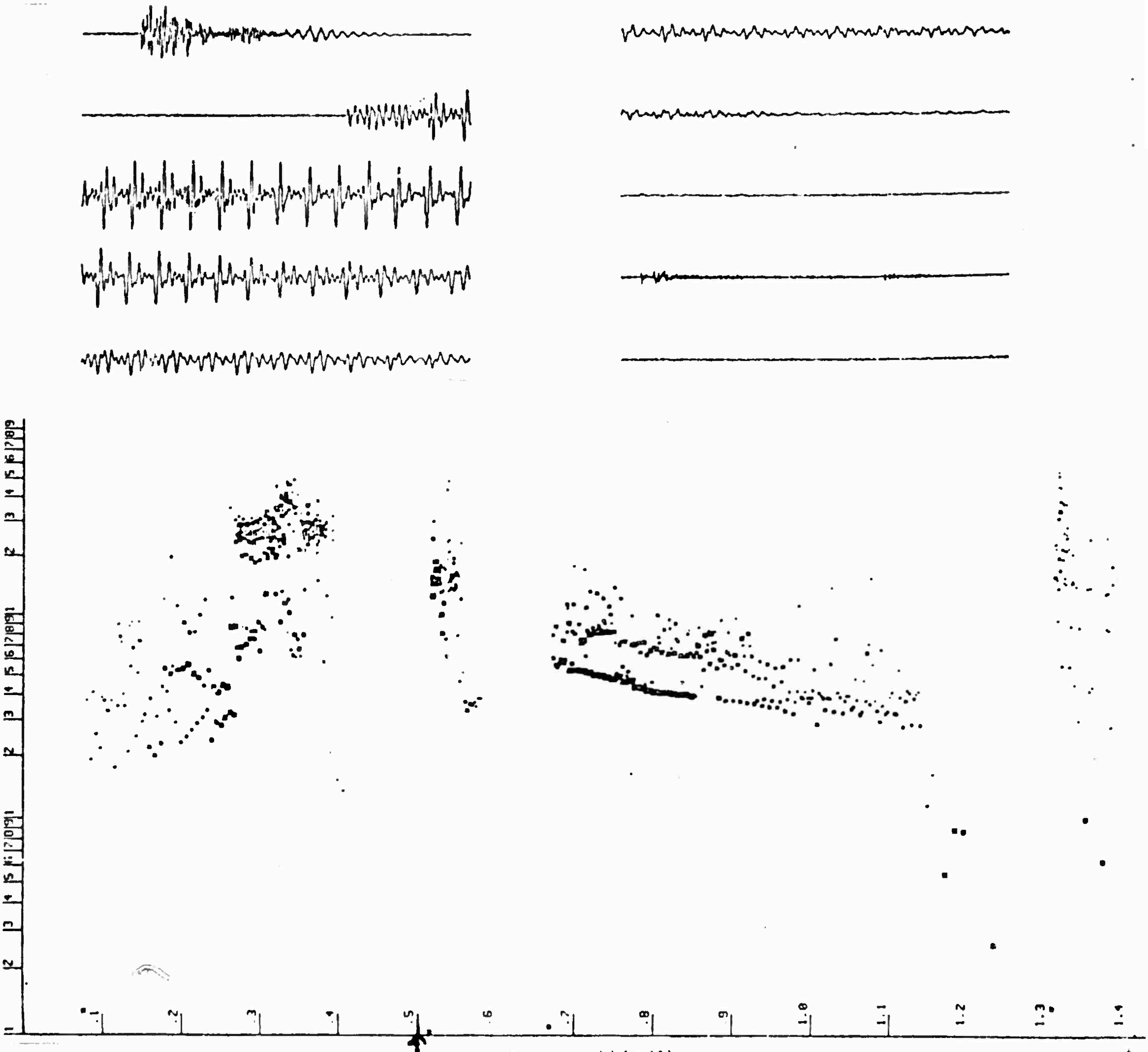


Figure 30: week old (σ,JA)

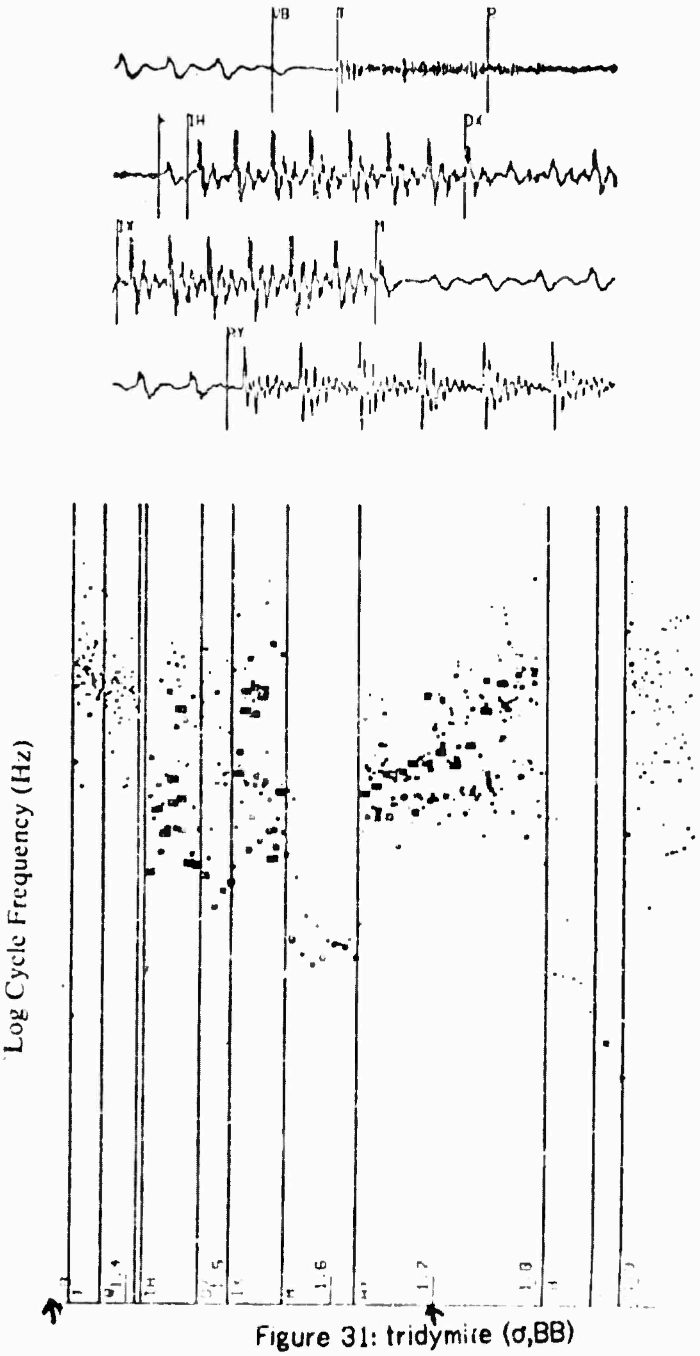


Figure 31: tridymite (σ,BB)

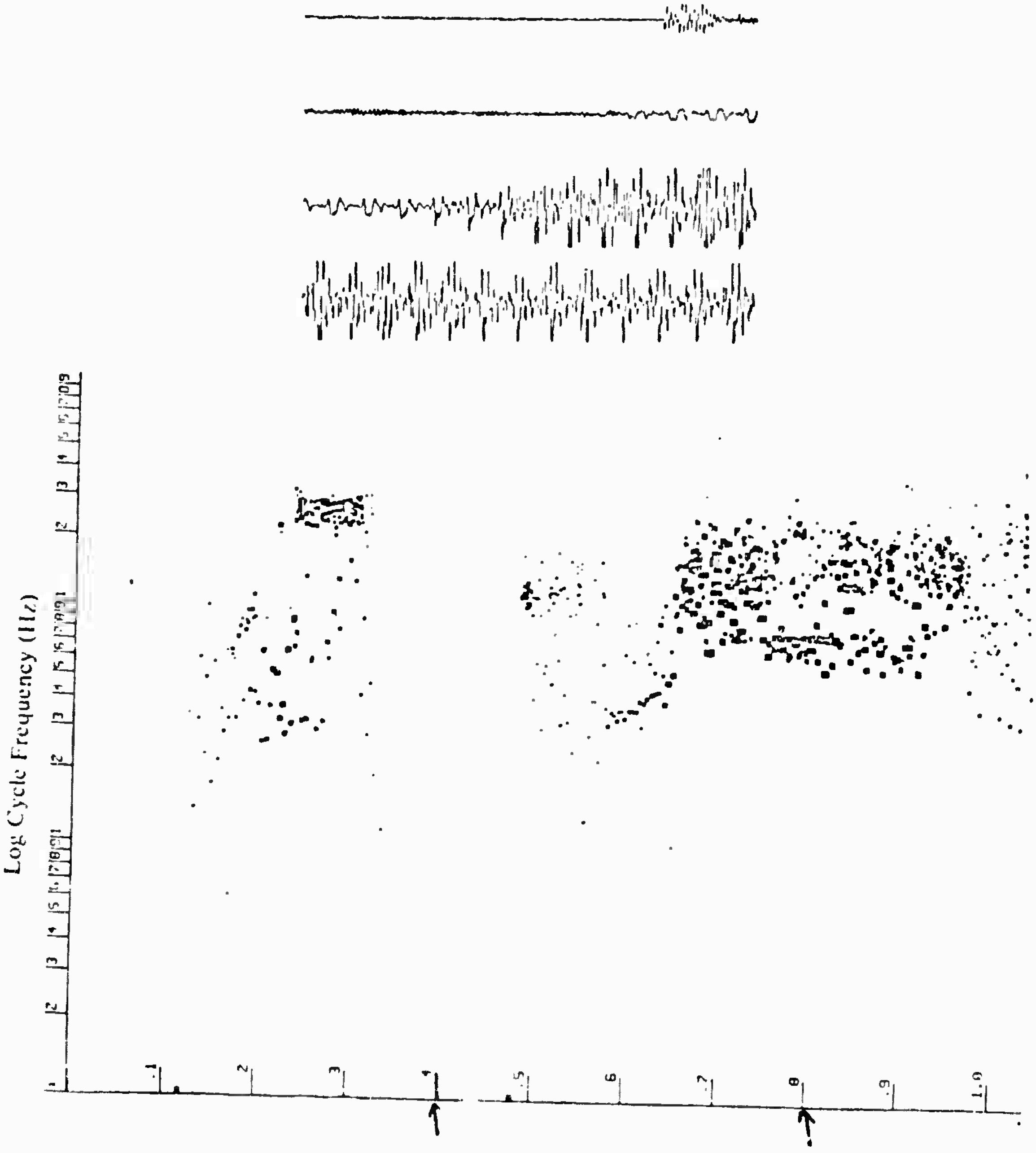


Figure 32 : week ran (σ,JA)

NASALS

Usually the nasals are characterized by large amplitude, low cycle frequency, and low microstructure. They are almost always very easy to spot visually and to segment automatically. They usually appear as a straight, almost horizontal, low cycle-frequency line. Beginning abruptly, they sustain acoustic consistency throughout their duration, and then usually terminate sharply.

The voicing regions which appear prior to voiced fricatives and stop consonants are acoustically very similar to nasals, except that voicing regions tend to be lower in amplitude. An example follows which illustrates both a typical /n/ and a voicing region preceding the /g/ in the utterance "egg nog" (Fig. 33,HN,♀).

The problem encountered by some frequency-domain analyses, of distinguishing nasals from liquids, is much less severe in the time-domain. The low cycle-frequency components of liquids almost always exceed 300 Hz whereas nasals tend to have lower cycle-frequencies just slightly below or above 200 Hz. In addition, liquids usually have higher cycle-frequency components with amplitudes larger than those of the higher cycle-frequency components sometimes occurring in nasals. An illustrative example demonstrates the distinc ion between the nasals /n/ and /m/, and the liquid /l/ (Fig. 34,LM,♀) in the utterance "phonemic labels".

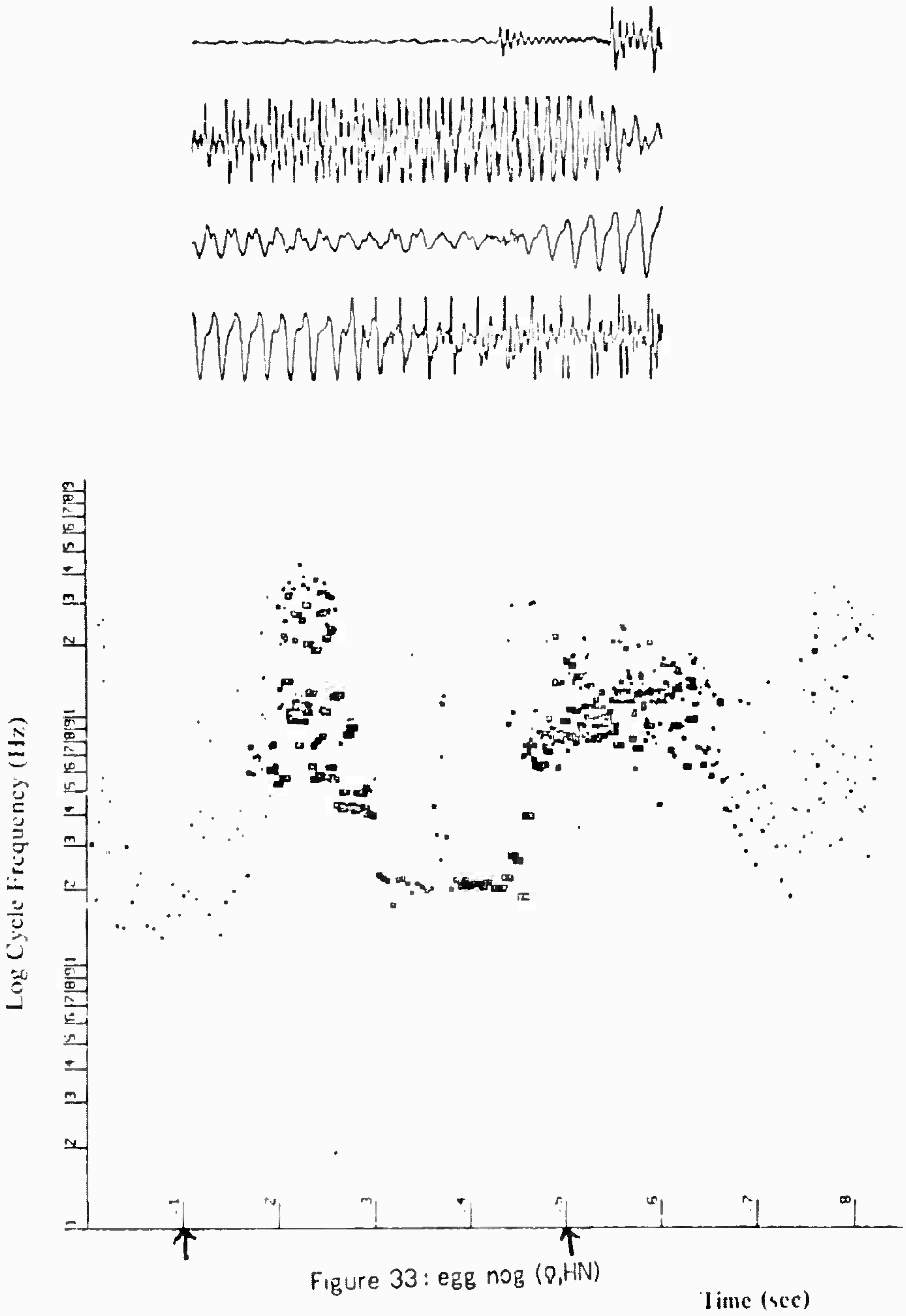


Figure 33: egg nog (♀,HN)

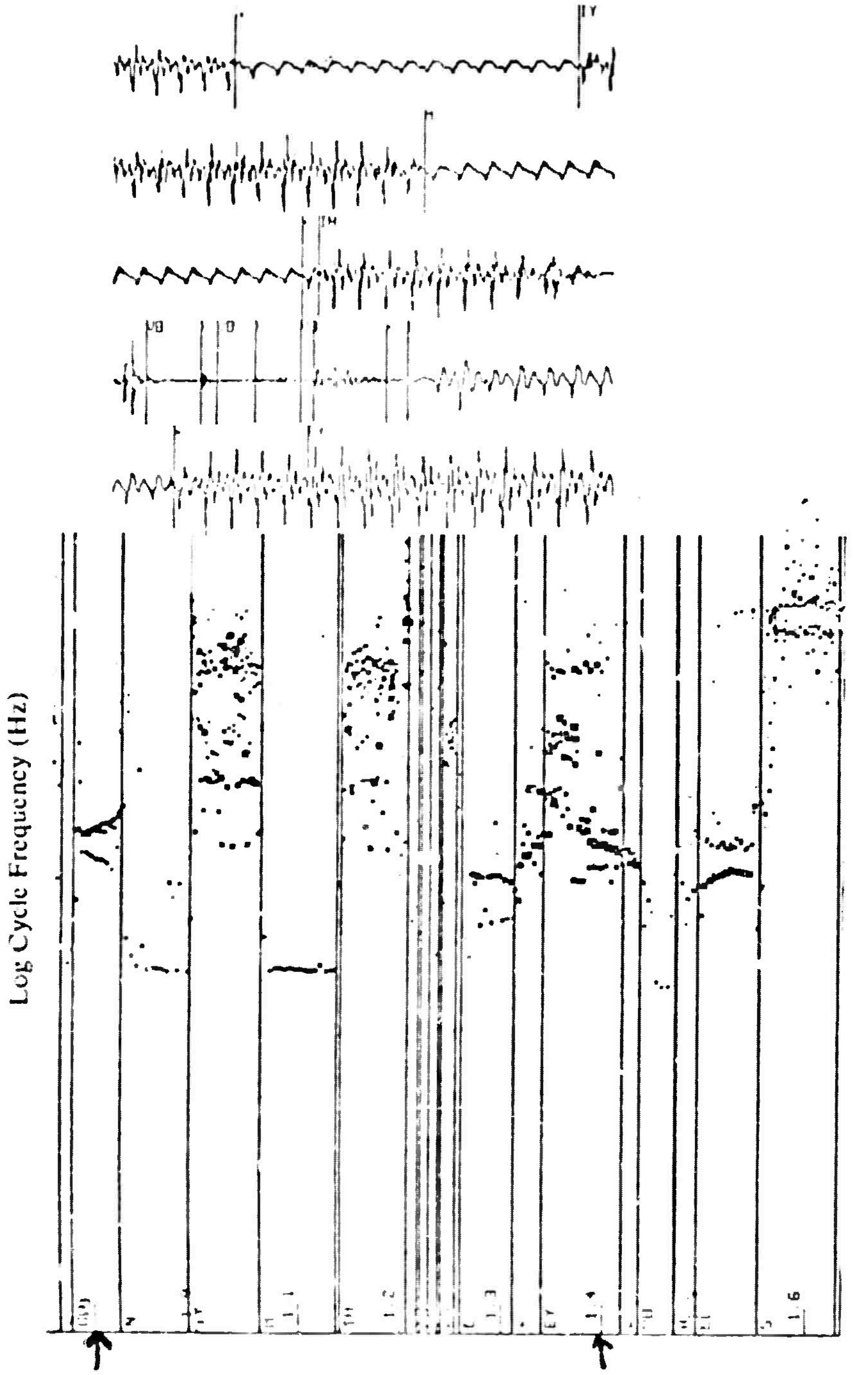


Figure 24: phonemic labels (9,LM)

DISCUSSION

One of the most striking aspects of speech as revealed by time-domain analyses, is its discrete nature. This fact is readily apparent in the waveform and LIP plots. Unfortunately, this quality of speech is often completely obscured in spectrographic displays, due to the inherent bandwidth limitation which averages acoustic observations over intervals, generally, of 5 msec, 10 msec, or more. This creates a visual (and digital) illusion of smooth gradual transitions from one acoustic state to another. However, precise temporal resolution shows that with the exception of liquids, the acoustic properties in the speech waveform change abruptly along one or more dimensions at phone boundaries. We also know that some phone types are characterized by an explicit temporal pattern of two or more distinct acoustic states. In many cases, additional redundancy for evidence of these boundaries is provided by the presence of transition cycles in and out of high cycle-frequency regions, stop previews, and stop preview transitions. Each of these events is also accompanied by sharp discontinuities in one or more time-domain parameters. These acoustic events clearly designate most inter-phone boundaries as well as delineating separate acoustic states internal to certain phones.

In the following table, we present statistics on the frequency of occurrence for the stop consonant and fricative features described in this chapter. In addition, we have found that we can use these time-domain features in conjunction with others which characterize the other phone classes, as the sole basis for automatic segmentation of the speech waveform. In Chapter III, the performance of our segmentation program is discussed in comparison with other existing segmentation programs, run on the same data set and generously made available by the speech community.

Frequency of Occurrence

for

Certain Time-Domain Speech Characteristics

On a set of 28 sentences included in the Lincoln Laboratory data set (described more fully in Chapter III), we obtained the following statistics for time-domain feature occurrences in stop consonants and fricatives:

	Percentage Occurrence
Stop Conscnants (n=135):	
Stop Preview Transition	15%
Stop Preview	26%
Voicebar and/or Pause	100%
Transition Cycle(s) (at end of Release-Aspiration)	68%
Fricatives (n=147):	
Beginning Transition Cycle(s)	66%
End Transition Cycle(s)	60%

Chapter III - SEGMENTATION OF CONTINUOUS SPEECH

INTRODUCTION

Speech segmentation usually implies the division of the speech waveform into a series of discrete acoustic states which are directly related to the phone string communicated by that waveform. This isolation process has become important both for understanding the basic acoustic characteristics of individual phones as well as comprising an essential step for phone identification in most speech recognition systems. In this chapter we present, first, a new segmentation philosophy and implementation, and secondly, the comparative results of this program with other segmentation programs presently available in the speech community.

THE SEGMENTATION PROGRAM

Our segmentation program is based solely on descriptions of the time-domain parameters which characterize a subset of the speech characteristics discussed in Chapter II. Subject to changes in these parameters, the segmenter transitions among eight acoustic states. These states are defined on the basis of our understanding of the temporal sequence of acoustic events characteristic of different phone classes. For example, based on the time-domain information about stop consonants and fricatives in Chapter II, each of these phone classes may be represented as a sequential pattern of acoustic events, some optional and some required.

In Fig. 35, the network describes the acoustic event pattern for both voiced and unvoiced stop consonants. The only absolutely required nodes in this network are the release-aspiration node and the preceding voicing and/or silence node(s). In the present segmentation program implementation, the stop preview and stop preview transition nodes are not represented because we found they yielded information redundant to that already available. Next, in Fig. 36, is the network representation for voiced and unvoiced fricatives. For unvoiced fricatives, the only required node is the frication node itself, and for voiced fricatives, the voicing node must precede this frication node.

Fig 35 - STOP CONSONANT PATTERN

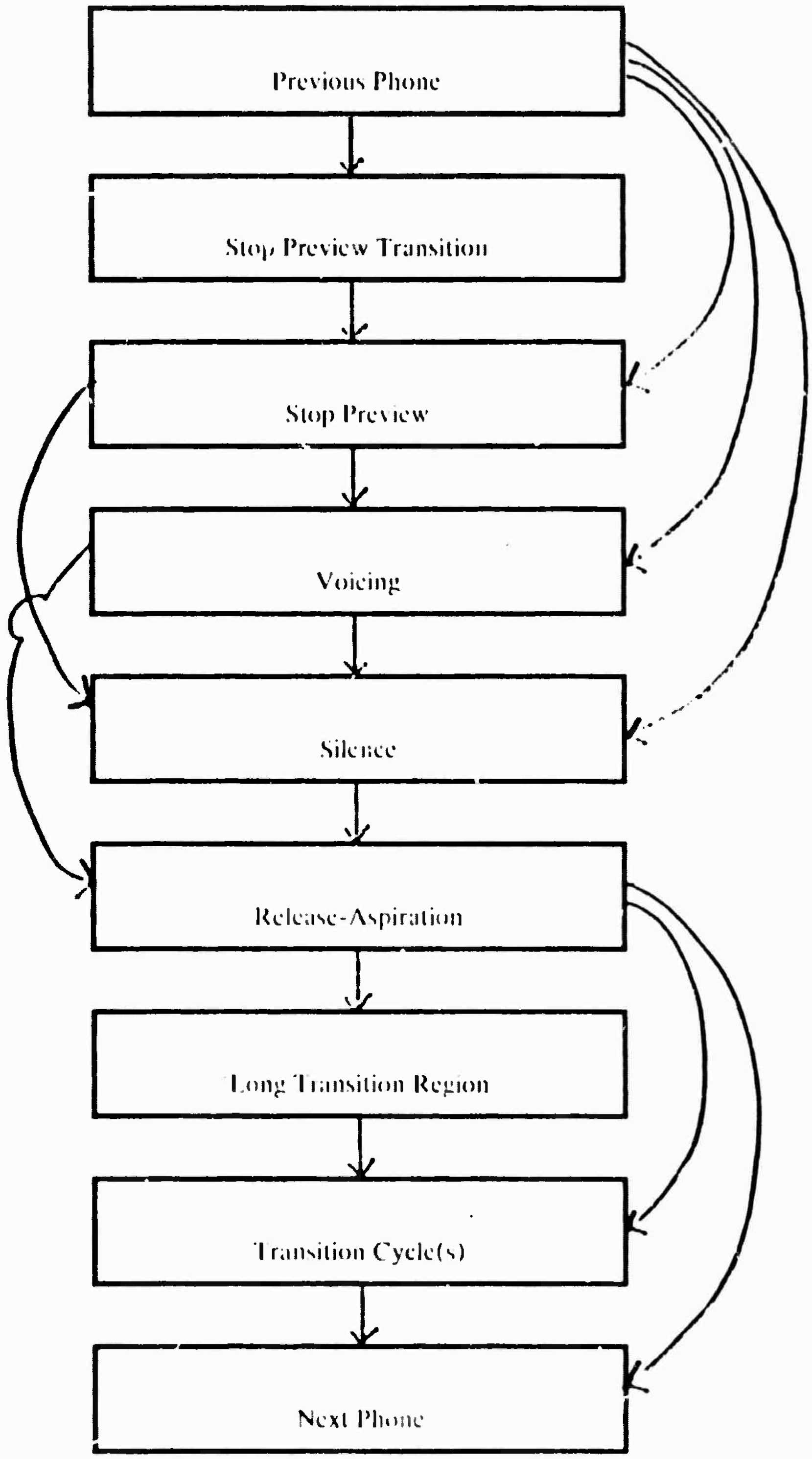
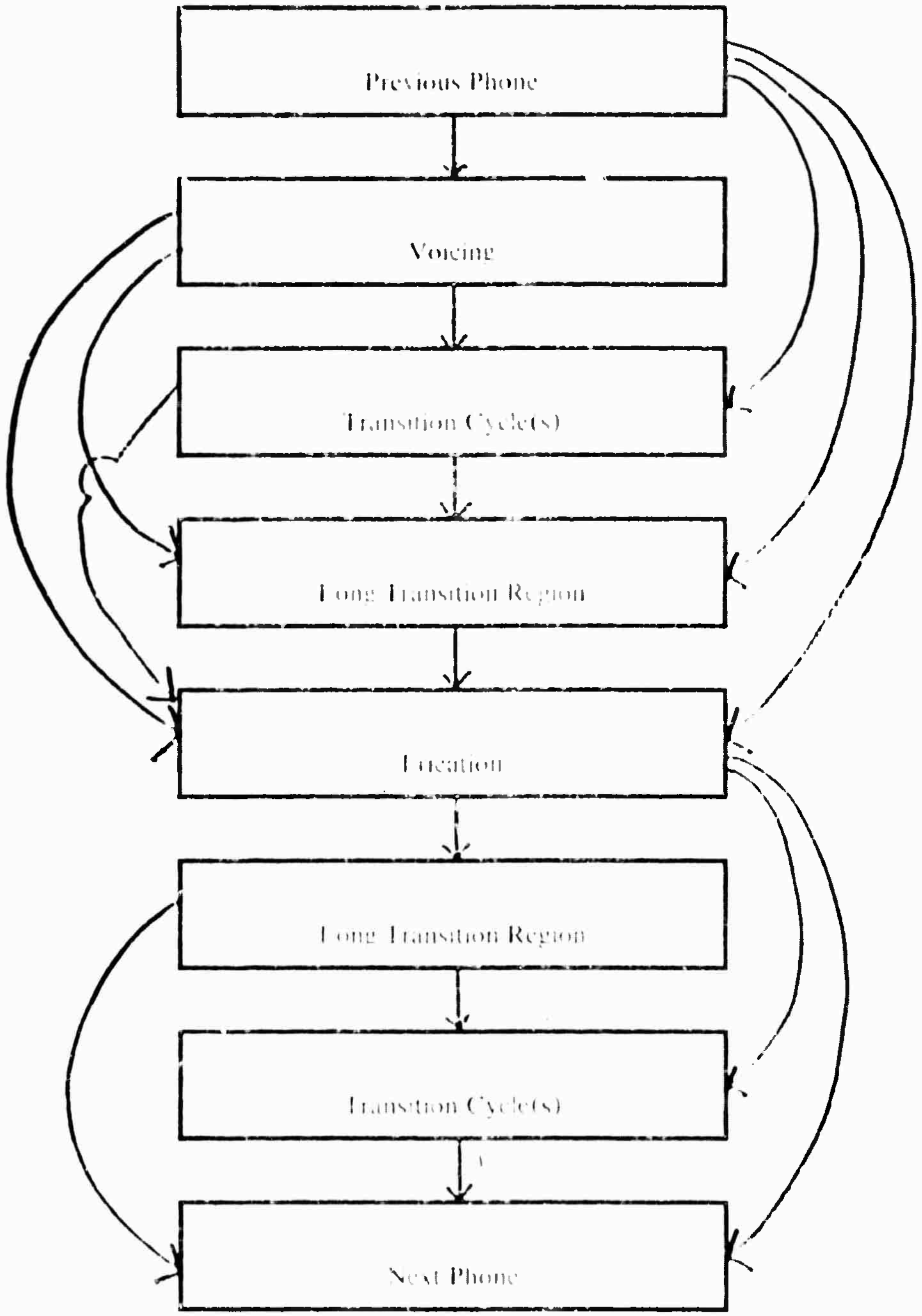


Fig. 36 FRICATIVE PATTERN



The eight acoustic states recognized by the program correspond very roughly to phone or sub-phone classes; namely, 1) silence (A), 2) unvoiced release-aspiration (B), 3) unvoiced frication (C), 4) nasal (D), 5) transition (E), 6) vowel (F), 7) voiced release-aspiration (G), and 8) voiced frication (H). Each of these states has already been described generally in Chapter II. Specific quantitative descriptions for each of these states as well as their transition characteristics, are incorporated explicitly under headings of the same name, in the segmentation program. Segment boundaries are placed when large changes of certain kinds are observed in the acoustic parameters. In the present implementation, we chiefly use the parameters of cycle-frequency and cycle peak amplitude (Amax), along with information about the acoustic state duration in terms of the number of cycles observed during a given acoustic state, and the identity of the past two acoustic states observed. In order to test the acoustic parameters of any given cycle, relative to its context, the acoustic parameters of the previous ten cycles and the following ten cycles, are available for comparison. By utilizing the information derived from the temporal phone patterns in conjunction with that available from cycle context comparisons, we allow for a great deal of individual cycle variability without triggering excessive numbers of extra or false segment markers. In addition, however, even very short duration low amplitude events such as fricative and stop consonant transition cycles are readily recognized because they conform to the known temporal phone patterns and bear known parameter relationships to their context. Redundancy of this sort is essential for segmentation decision reliability.

The program itself is quite simply organized Each cycle of the waveform is evaluated in turn, to determine if it belongs to the present acoustic state, or if it marks the beginning of a new acoustic state. A certain set of state-dependent tests is applied to the parameters of each cycle. In Part 1 of the program, only those tests associated with the present acoustic state are applied to the parameters of a new cycle (each utterance is abitrarily initialized to the silence state A).

In Fig. 37, all the acoustic state tests are fully described. Here CF_t, A_t, and D_t, respectively, designate, for the present cycle (present time = t), the paramcter values of cycle-frequency, cycle peak amplitude, and the present acoustic state duration, up to but not including the present cycle.

For example, if the present acoustic state is the silence state A, then only two tests are applied to a given cycle. These tests are applied in the order indicated. The first says that if the cycle-frequency of this cycle is greater than 120 Hz, and either the present cycle or the next cycle has a cycle-frequency greater than 600 Hz, then transition to state B. The second test says that if the cycle-frequency values of the present cycle and the next cycle both exceed 200 Hz, then transition to state E. If the conditions of any of these tests are met, then a segment marker is inserted, a new acoustic state is recognized, and the program then commences to evaluate the next cycle in the same fashion. Whenever none of the tests is met for transitioning out the present acoustic state, and the present acoustic state has already been observed for two or more cycles, then three more tests are applied to this cycle. These three tests comprise all of Part II of the program. These three tests are described in Fig. 38.

In some cases, these preceding tests invoke the additional tests TT, TP, TN, and TS. The first of these, TT, tests to see if a return should be made to an acoustic state previously encountered. Here "oldstate" refers to the acoustic state immediately preceding the present one, and "oldoldstate" refers to the state just prior to the oldstate. The other tests, TP, TN, and TS, are general tests for signal periodicity, nasality, and frication, respectively. All of these tests are also described in Fig. 38. If all the necessary tests in Part I and Part II are applied, and none are met, this newly examined cycle is assumed to belong to the present acoustic state, and testing then commences for the next cycle.

PART I

STATES	TESTS
A	1) if $(CF_t > 120) \wedge (\max(CF_t, CF_{t+1}) > 600 \rightarrow$ **B**
(≈silence)	2) if $\min(CF_t, CF_{t+1}) > 200 \rightarrow$ **E**
B	1) if $CF_t < 120 \rightarrow$ **A**
(≈unvoiced release-aspiration)	2) if $(A_t \geq 100) \wedge (TP) \rightarrow$ **G**
	3) if $(CF_t < 200) \rightarrow$ **E**
	4) if $\max(CF_t, CF_{t-1}) < 600 \rightarrow$ **E**
C	
	1) if $CF_t < 120 \rightarrow$ **A**
(≈unvoiced frication)	2) if $(A_t \geq 100) \wedge (TP) \rightarrow$ **H**
D	1) if $CF_t < 100 \rightarrow$ **A**
(≈nasal)	2) if $(CF_t > 650) \wedge (CF_{t+1} > 400) \wedge ((CF_{t+2} > 400) \vee (CF_{t+3} > 400)) \rightarrow$ **B**
	3) if $(CF_t > 350) \wedge (CF_{t+1} > 250) \wedge (A_{t+2} > 350) \vee (A_{t+3} > 350)$ **F**

Figure 37

STATES	TESTS
E (: transition region)	1) if $(CF_{t-1} > 350) \wedge (D_t > 1) \rightarrow F$
F (: vowel)	1) if $(CF_t \geq 1000) \wedge$ (2 or more of $(CF_{t+1}, CF_{t+2}, CF_{t+3}, CF_{t+4}) \geq 1000 \rightarrow TT$
G (: voiced release-aspiration)	1) if $(CF_t < 200) \wedge (CF_t \geq 120) \rightarrow E$ 2) if $(CF_t < 120) \rightarrow A$ 3) if $(A_t < 100) \wedge (\neg TP) \wedge (\max(A_{t-1}, A_{t-2}) < 100) \rightarrow B$
H (: voiced frication)	1) if $(CF_t < 200) \wedge (CF_t > 120) \rightarrow E$ 2) if $CF_t < 120 \rightarrow A$ 3) if $(A_t < 100) \wedge (\neg TP) \wedge (\max(A_{t-1}, A_{t-2}) < 100) \rightarrow C$ 4) if $(\max(CF_t, CF_{t+1}) < 1000) \wedge ((CF_{t+2}) + (CF_{t+3}) < 3000) \rightarrow E$

Figure 37 (Cont.)

PART II

For all states with duration > 1:

1) if $(\neg A) \wedge (\neg D) \wedge (CF_t < 120) \rightarrow A$

2) if $(\neg D) \wedge (CF_t < 350) \wedge (CF_t > 100) \wedge (\max(A_t, A_{t+1}) > 120) \rightarrow TN$

3) if $(\neg A) \wedge (\neg B) \wedge (\neg C) \wedge (\neg G) \wedge (\neg H) \wedge (CF_t > 1000) \rightarrow TS$

TESTS

TT

(test oldstate)

1) if (oldstate=E) $\wedge$ (oldoldstate=B C G H) $\rightarrow$ oldoldstate

2) $\rightarrow$ C

TP

(periodicity check)

Over the present and next 9 cycles, we measure the amplitude range observed and the respective ratios of high vs. low (< 100) amplitude cycles and high vs. low (< 1000) cycle-frequency cycles This test is "true" for regions of amplitude and cycle-frequency periodicity.

TN

(nasality check)

If 2 or more of the next 4 cycles have cycle-frequency < 400 $\rightarrow$ D

TS

(frication check)

Over the following 9 cycles, this test checks to see there are no low frequency cycles (< 1000) and no more than 4 cycles between 1000 and 2000 Hz, in order to designate this region as containing frication.

Figure 38

TESTS

1) if F + E(dur. < 10 msec) → F

2) if H(dur. > 20 msec) + C(dur. < 20 msec) + H(dur. > 20 msec) → H

3) if C(dur. > 20 msec) + H(dur. < 20 msec) + C(dur. > 20 msec) → C

4) if G(dur. > 20 msec) + B(dur < 20 msec.) + G(dur. > 20 msec) → G

5) if B(dur. > 20 msec) + G(dur. < 20 msec) + B(dur. > 20 msec) → B

Figure 39

At the end of an utterance, the output of this program consists of a series of times at which segment markers were inserted with their respective acoustic state designations. This segmentation output is then run through an editing program which concatenates certain kinds of segments, and deletes very brief segments of noise-like regions.

This editing program, described in Fig. 39, only alters certain boundaries of acoustic segments which are less than 20 msec in duration. If the acoustic state F (of any duration) is followed by state E, where E has a duration of less than 10 msec, than the segment marker between states F and E is removed, and the combined region is classified as state F. If an unvoiced frication segment of less than 20 msec duration, is both preceded and followed by voiced frication regions of duration equal to or greater than 20 msec, then the unvoiced frication segment boundaries are removed, and the previously defined sequence of the three segments (voiced, unvoiced, and voiced frication) is classified as a single voiced frication segment. Such merging also occurs if a short voiced frication segment is surrounded by longer unvoiced frication segments, which then prevail. In the same way, merging is performed for successive voiced-unvoiced release-aspiration segments of stop consonants.

COMPARATIVE SEGMENTATION RESULTS

In July, 1973, a Segmentation and Labeling Workshop was held at the Computer Science Department of Carnegie-Mellon University, Pittsburgh, Pa. In preparation for this workshop, certain of the speech groups contracted by ARPA (Advanced Research Projects Agency) submitted a set of continuous speech utterances typical of the input to their speech understanding systems under development. At Lincoln Laboratory, a subset of these, 31 utterances in all, were chosen, prepared, and digitized at 10 kHz and 20 kHz. Analog and digital tapes of these were then made available to all groups in the speech community at large, who wished to submit their results for segmentation and/or labeling procedures on this set of data. At the workshop, these results were compared and discussed.

This section describes the results of using time-domain analyses for automatic segmentation of continuous speech. As previously described, our program looks for specific temporal event patterns and discontinuities in various time-domain parameters, as indicators of phonetic boundaries. The output of this program has been compared with the output of four other automatic segmentation programs presented at this speech worksnop (or the improved results subsequently submitted), on a set of five utterances, one utterance for each of five speakers (4male, 1 female).

These utterances follow:

1) Do any samples contain tridymite?

2) Count where type equals linear equations and runtime less than five six.

3) I want to do phonemic labeling on sentence six.

4) Display the phonemic labels above the spectrogram.

5) Do you have any rectangular cylinders left?

In Appendix A, there are two different visual displays for each of these utterances. These are 1) the LIP plots, and 2) the spectrograms (prepared for the workshop).

This comparison test is a particularly strenuous test of robustness because each of the five programs segment, without any speaker training, utterances spoken by five different speakers recorded under different environmental conditions. Since the segmentation programs of all the groups represented are still being actively developed (including this author's program), none of the segmentation results presented here should be, in any sense, construed as optimized or finalized. Despite the preliminary nature of all of these results, we feel that since each of these programs has been run on the same data set, a comparison of these is the best and most appropriate test presently available. In the discussion which follows, our program results are designated by "TDS" (time-domain segmentation). The segmentation results of the other groups are designated as "B", "C", "D", and "E", respectively.

As the first step, we carefully hand-segmented each utterance with the assistance of waveforms, spectrograms, LIP plots, and records of cycle-by-cycle time-domain parameters. With few exceptions, our hand segmentation agreed closely with others prepared for the workshop. In this hand segmentation, all boundaries between acoustically distinct segments were marked. Then the output of each automatic-segmentation was compared with this hand segmentation. Certain conventions for doing this were established.

Primary boundaries were designated as those comprising the minimal set of boundaries considered essential for a basic segmentation of the acoustic waveform. First it was assumed that one primary boundary exists between every two successive phones. In those instances where sharp acoustic discontinuities do not occur between two phones (as frequently occurs at the beginnings and ends of liquids), a best approximation of such a boundary was made. For stop consonants, two primary boundaries were considered necessary, one designating the pause region and another at the onset of the release-aspiration region. A primary boundary was also required to designate the onset of voicing if it lasted for more than 20 msec prior to the high cycle-frequency region of a fricative or stop consonant. Secondary boundaries mark regions which are acoustically distinct but are not necessarily related phonetically to the speech stream. Transitional segments between two successive phones are marked by secondary boundaries. In addition, where there is a long transition in or out of a phone such that the acoustic characteristics of this transition differ substantially from the acoustic characteristics prototypic of that phone, a secondary boundary demarcates this region.

The aim in comparing each automatic segmentation to the hand segmentation is to evaluate how many primary boundaries were found and by how much time they were displaced from the hand segmented boundaries. Where a transitional segment occurs between two phones in the hand segmentation, the primary boundary between those phones is considered to be that boundary, as provided by a given automatic segmentation procedure, which is closest to one of the two boundaries provided by the hand segmentation. An example is shown in the following excerpt of a single phone, /t/, in a multiple segmentation comparison plot from the first utterance "Do any samples contain tridymite?". In the multiple segmentation comparison plot which follows in Fig. 40, the segmentations produced by the different programs (TDS, B, C, D, and E), appear on successive lines under the waveform they segment. Up-arrows denote segmentation markers. In the hand segmentation shown above the waveform of this /t/ phone, a transition segment is marked at the end of the release-aspiration. This transition segment is marked with a double-headed arrow). In this example, we denote the beginning marker of this transition segment with "I", and the end marker as "II". The multiple segmentation comparison plot shows that TDS places a marker closest to II. Further, we see that B has a marker closest to II, C to I, D to I, and E to II.

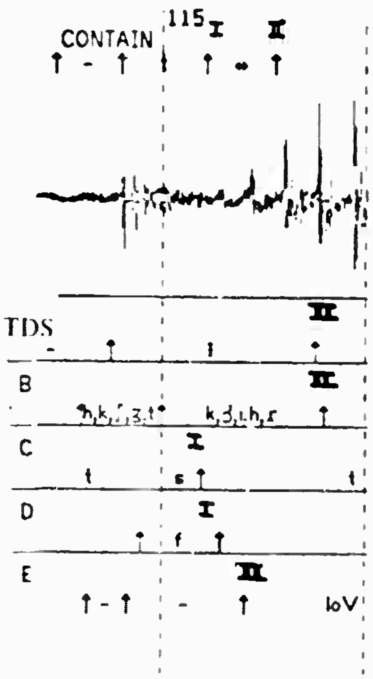


Figure 40: transition segment following /t/ in "contain" (σ,BB)

For each segmentation program then, the deviation in time difference is computed from the marker it provides, to the closest marker provided by the hand segmentation, as shown above.

Similarly, if a transition segment is marked in an automatic segmentation, whichever of the two markers, is the closer to the hand segmentation marker, that marker is designated as the primary boundary, and the other as a secondary boundary. In computing the absolute deviation of automatic segmentation boundaries, only the primary boundaries are considered. When an automatic segmentation provides for more than one primary boundary where only one boundary exists in the hand segmentation. the unmatched marker(s) are considered to be "extra". Automatic segmentations which insert no boundaries within 35 msec of primary boundaries in the hand segmentation, are considered to have missing boundaries. However, if an unmatched boundary is inserted more than 35 msec from a missing boundary, it is not counted as being an extra boundary.

We have another example (Fig. 41) from the first utterance "Do any samples contain tridymite?". We compare each of the automatic segmentations with the hand segmentation, for the first three phones of the word "tridymite". Above each of the automatic segmentation markers in the comparison plot, we have labeled the marker as "p" for primary, "s" for secondary, or "e" for extra. Areas where a primary boundary is missing, are labeled with an "m". Note that TDS and E designate transitional segments in different ways; TDS uses a double-headed arrow segment label whereas E uses left-barbed and right-barbed half arrows for marking the transition segment boundaries.

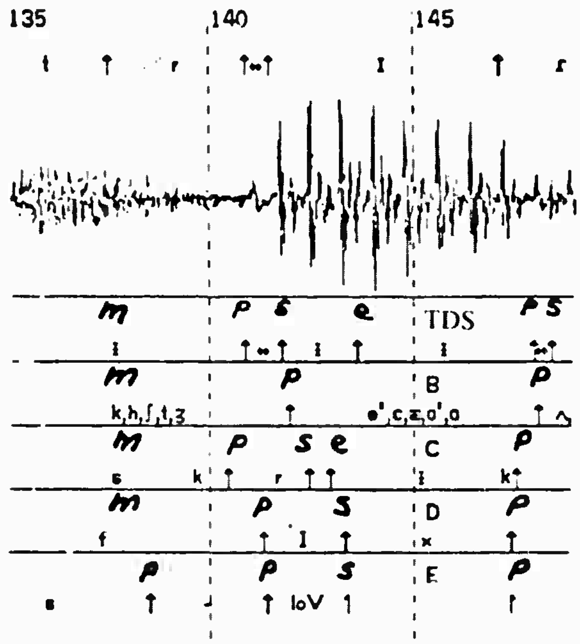


Figure 41: comparison of some segments ir "tridymite" (σ,BB)

An overall tally was computed for each automatic segmentation program, for each utterance, by separately totalling the number of primary boundaries, secondary boundaries, extra boundaries, and missing boundaries. The primary boundaries found by each program were further analyzed within each utterance. These automatically derived primary boundaries, each characterized by its absolute deviation in time from the hand-segmented boundaries, were separated according to general phone classes. For example, the primary boundaries which designated the start of vowels were grouped together, as were those for pauses, stop consonants, liquids, fricatives, nasals, and voicing regions which preceed the voiced fricatives and stop consonants. Then for each class, of phones, an average was computed of the absolute deviations in time from the hand-segmentation pı boundaries.

A summary table (Fig. 42) combines the results from all the utterances combined, for each segmentation program individually. The total number of phones found in each class, is designated in parentheses following the absolute time deviations (expressed in milliseconds) of the primary boundaries in each class. In addition, we also have computed the accuracy for each segmentation program in finding the end primary boundaries of fricatives and stop consonants. In Appendix B, we present separate analyses for each of the five utterances.

ABSOLUTE DEVIATION (MSEC) OF PRIMARY BOUNDARIES BY PHONE CLASS

PROGRAMS	PAUSES	STOPS	VOWELS	LIQUIDS	FRICATIVES	NASALS	VOICING
TDS	2.3(24)	1.3(35)	6.0(49)	7.0(18)	4.7(23)	1.6(25)	2.0(22)
B	11.6(13)	8.6(26)	11.4(55)	7.1(21)	10.8(21)	11.7(22)	11.2(18)
C	8.3(9)	9.1(13)	11.7(43)	8.3(8)	13.5(21)	8.2(18)	6.0(14)
D•	7.0(14)	4.5(20)	9.1(40)	10.2(14)	11.7(13)	7.8(17)	10.0(12)
E	10.7(18)	11.0(20)	12.7(49)	12.5(20)	14.2(19)	11.6(17)	5.6(17)

PRIMARY END BOUNDARIES			SECONDARY BOUNDARIES
PROGRAMS	STOPS	FRICATIVES	# FOUND
TDS	5.1(24)	5.0(22)	46
B	9.0(25)	11.6(22)	9
C	8.1(19)	12.9(16)	4
D	8.1(18)	9.6(14)	6
E	11.8(24)	14.1(23)	9

PROGRAMS	PRIMARY BOUNDARIES FOUND		AVERAGE ABSOLUTE DEVIATION (msec)
	# found	percentage (total=216)	
TDS	194	91%	3.5
B	176	81%	10.6
C	126	58%	10.1
D•	130	86%	8.8
E	160	74%	11.5

PROGRAMS	EXTRA BOUNDARIES FOUND		MISSED PRIMARY BOUNDARIES		MISSED PLUS EXTRAS	
	# found	percentage (total=216)	# missed	percentage	# total	percentage
TDS	38	18%	20	9%	58	27%
B	43	20%	40	19%	83	38%
C	9	4%	90	42%	99	46%
D	51	34%	22	14%	73	48%
E	32	15%	56	26%	88	41%

PROGRAMS	PERCENTAGE STOPS FOUND (total=36)	PERCENTAGE FRICATIVES FOUND (total=24)
TDS	97%	96%
B	72%	88%
C	39%	88%
D	83%	81%
E	61%	79%

There are several important aspects to consider while comparing the results of these five different automatic segmentation programs. Segmentation programs are designed to produce segment markers when they detect changes from one acoustic state to another. The amount of acoustic change necessary for triggering detection of a new acoustic state is generally determined by one or more thresholds.

Experience has shown that with many segmentation programs, systematically varying the threshold value(s) causes results which range all the way from leaving out very few real boundaries but inserting many extra boundaries, to missing many of the real boundaries but adding very few extra ones. Depending on how a given program is to be used, these thresholds must be set to minimize segmentation errors of certain kinds. For example, a segmentation program might be designed chiefly to locate certain easily recognizable acoustic states ("islands of reliability"). In this case, very conservative thresholds might be used to obtain a segmentation which accurately locates most of the best defined boundaries, misses many of the less distinct boundaries, and inserts very few if any extra boundaries. However, another segmentation program which is designed to locate as many real acoustic boundaries as possible, would probably set lower thresholds for locating new acoustic states. Consequently this might produce results where most of the true acoustic boundaries are detected (though less distinct boundaries might not be so accurately determined in time as sharper boundaries) and therefore very few boundaries are missing, but a great many extra boundaries might also be inserted. In short, there exists a trade-off between the number of missing boundaries and the number of extra boundaries.

Previously, we have described the kinds of sharp discontinuities which exist at most phone boundaries, and certain temporal patterns of acoustic changes characteristic of fricatives and stop consonants. Our segmentation program incorporates a large subset of these features. In particular, we aimed for detecting most of the primary boundaries and locating them as accurately as possible, especially the stop consonants for which accurate location is most crucial.

As indicated in the preceding table, the TDS program detected more boundaries (91%) than the other programs, and located these 250% - 325% more accurately when the absolute deviations for all the phones were averaged together. Only C had fewer extra boundaries (4%) than

TDS (18%), but on the other hand, it also missed 42% of the primary boundaries as compared to 9% for TDS. This is an example of the missing-extra boundary trade-off. However, TDS was also able to find the greatest number of fricatives, stop consonants, and secondary boundaries, as well as pauses, nasals, and voicing regions.

Detecting the release-aspiration segments of stop consonants is very important because they last for such a short time. The average duration for the release-aspiration region of stop consonants in all five utterances was 25.1 msec with several shorter than 5 msec. Fricatives, by contrast, averaged 79.9 msec in length. TDS typically located the start of release-aspiration with an error of 1.3 msec. As compared with the other segmentation programs, the TDS boundaries were about 350% to 850% more accurate. Although the absolute magnitude of their errors was small, from 4.5 msec to 11.0 msec, these errors were large relative to the release-aspiration durations. After these errors were further convolved with several milliseconds more of error in determining the stop consonant end boundaries, it is no surprise that many stop consonants were missed altogether by the other programs. For the other segmentation programs, the sum of beginning boundary and end boundary errors typically averaged 2/3 of the release-aspiration duration. This means that if these segments are used either for providing acoustic templates or training for acoustic recognition programs, or are used for comparison against other templates, good recognition results are doubtful. As for other classes of phones, the average error of TDS in locating fricatives was 4.7 msec, for pauses 2.3 msec, for nasals 1.6 msec, and for voicing regions 2.0 msec. Although nasals and voicing regions are usually acoustic steady states, they begin abruptly, are well characterized throughout, and end abruptly. As can be seen from Fig. 42, although programs B and E found several more liquids and/or vowels than TDS, the TDS program, at its worst overall, was as accurate or more accurate than all of the other programs in locating liquid and vowel primary boundaries.

Another issue which arises is the robustness or consistency of segmentation procedures across different speakers and recording conditions. For purposes of comparison, tables of absolute deviations by phone classes are presented for each utterance, in Appendix B, as previously mentioned. consistent. We observe, for example, that for utterance #3, C accumulated the lowest total of missing plus extra boundaries, only 6 in all, whereas in utterance #4, they accumulated the highest total of missing plus extra boundaries, 14 in all (33 missing, 3 extra). It should be noted though, that utterance #4 was the only utterance spoken by a female (LM). Since male voices have been used almost exclusively for acoustic speech research, the male speech characteristics incorporated into some segmentation/labeling programs may preclude good performance on female voices. If the fourth utterance is not included for C, their score at finding primary boundaries for all the utterances combined increases from 58% to 66%, and for detecting stop consonants, it increases from 39% to 41%. Segmentation results for D were not available either for the last half of the third utterance or the entire fourth utterance. Therefore their percentage scores are computed proportionately to the material for which their segmentation results were available. Hand segmentation shows the material which D completely segmented, to contain a total of 152 primary boundaries, including 16 fricatives and 24 stop consonants

THE PROBLEM

Our study of allophones is prompted quite simply by the fact that the acoustic manifestation of any phone is, to varying degrees, a function of the context or acoustic environment in which it occurs. As a consequence, we must understand these coarticulation phenomena in order to discriminate well between phones of the same class, among the stop consonants, for example. We use the term "allophone" to designate a phone embedded in a certain kind of environment, with respect to post- and/or pre-context. For example, the allophones of /k/ include those which are nasalized, retroflexed, rounded, and so forth. In general, it has been recognized that the context following a given phone, exerts a greater influence on that phone's acoustic manifestation than does its preceding context.

An extreme example of contrast in the acoustic characteristics of a phone, due to context differences, is demonstrated by a comparison of Figs. 43 and 44. Here the two allophones to be compared are the unrounded /k/ and the rounded /k/, in the words "king" and "queen", respectively. The utterance of Fig. 43 is "Pawn to king four", and the utterance in Fig. 44 is "Pawn to queen four". In these figures, it is apparent that the average cycle-frequency of the release-aspiration of the /k/ in "queen" is substantially lower than that of the /k/ in "king". This is due to the rounding of the lips which effectively lengthen the vocal tract, thereby lowering the frequencies emitted. In addition, we know that for both /k/ and /g/, in general, the precise place of articulation varies somewhat as a consequence of coarticulatory factors. Although it comes as no surprise that different allophones of the same phone have different acoustic characteristics, it does create certain problems. It means that the ability to recognize a given phone in one context is not generalizable to recognition of that phone in other contexts. In perception experiments, Schatz [S2] has demonstrated that context is vital for proper identification of voiceless stop consonants. Given almost any parameterization, a great deal of overlap between similar phones is unavoidable. The intra-phone differences, (that is, the differences between allophones of a given phone) are nearly as great as the inter-phone differences. Therefore studies are required on the acoustic variations of allophones. In the past, this work has been mostly concerned with frequency-domain examination of vowel allophones [S4]. And much of this has been centered on analyses of vowel formant trajectories [S5].

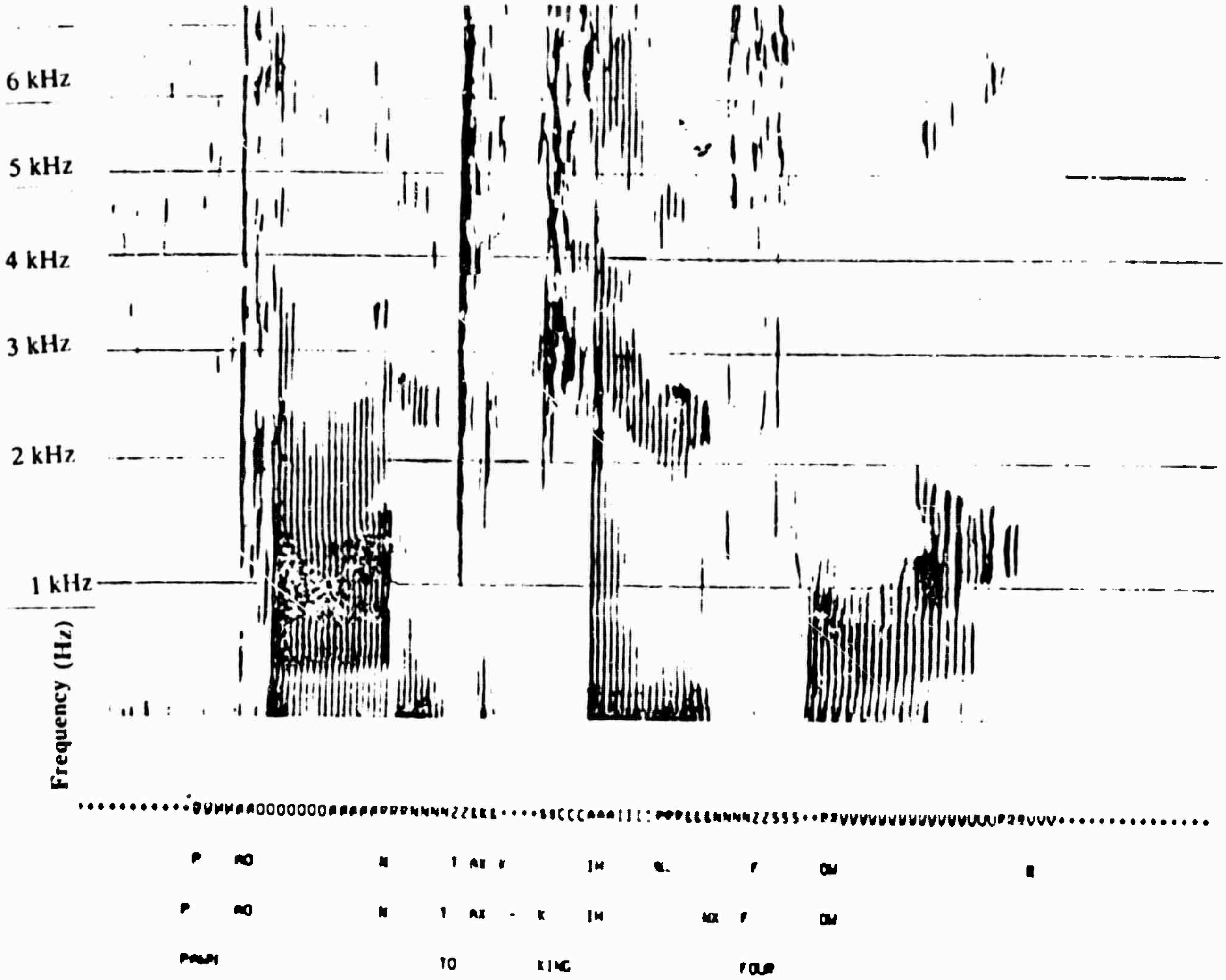
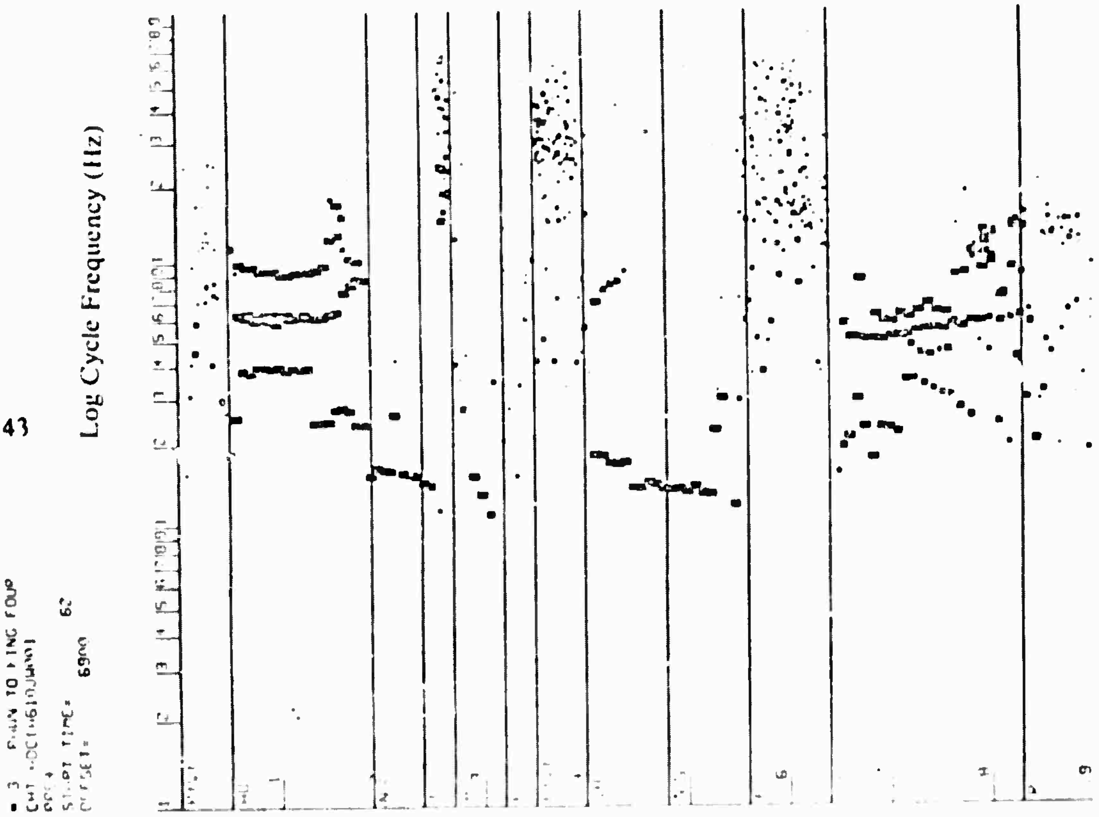

6 kHz
5 kHz
4 kHz
3 kHz
2 kHz
1 kHz
Frequency (Hz)
P AO N T AX K IH NG F OW R
P AO N T AX - K IH NX F OW
PAWN TO KING FOUR

Fig. 43


Log Cycle Frequency (Hz)
3 PAWN TO KING FOUR
START TIME= 62
OFFSET= 6900

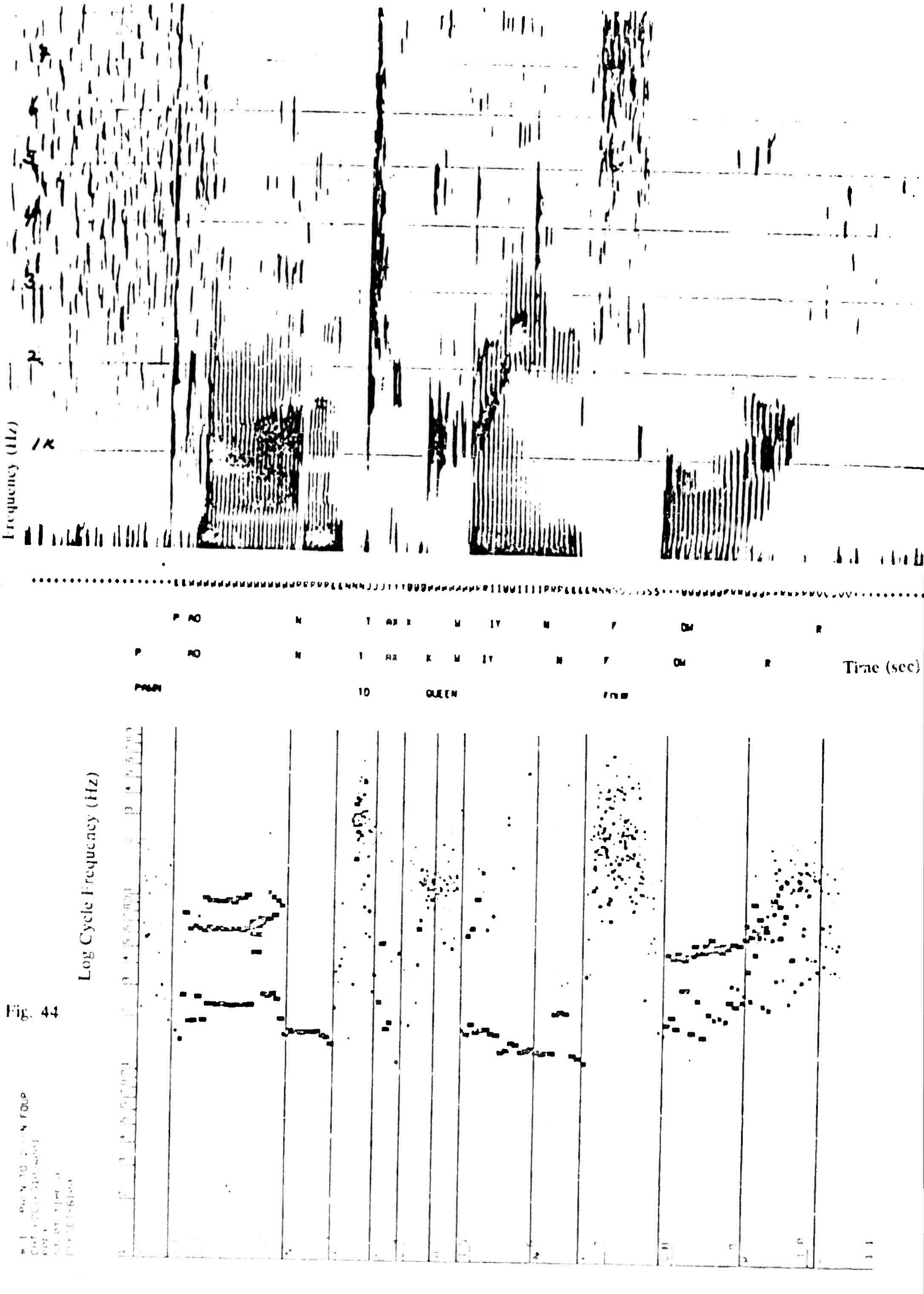

Frequency (Hz)
7
6
5
4
3
2
1K
P AO N T AX K W IY N F OW R
P AO N T AX K W IY N F OW R
PAWN
TO
QUEEN
FOUR
Time (sec)
Log Cycle Frequency (Hz)

Fig. 44

AN ALLOPHONE EXPERIMENT

Since time-domain techniques are particularly suitable for the acoustic characterization of fricatives and stop consonants, we have chosen to study the allophones of these. Shoup has generously made available to us, audio tape recordings of 3 speakers, all linguists, (2 male, 1 female), each reading 228 utterances encompassing both the common and rare allophones of the fricatives and stop consonants in general English. Details on the specific allophone designations, including phonetic transcriptions for all of these, have previously been published by Shoup [S3]. Each of these utterances is in citation form; e.g. "no foe", "tube moves", with the allophone of interest embedded in a suitable context. The allophones enumerated for the fricatives include nasalization, retroflexion, dentalization, and rounding, and for the stop consonants include aspiration, nasalization, retroflexion, palatalization, dentalization, and rounding, as well as the minimally released form. For both phone classes, the unmodified allophone is also included; this generally occurs in a neutral context of high front vowels. Not all allophones are possible for all phones. For the stop consonants, dentalization is restricted to /t/ and /d/, while palatalization is restricted to /g/ and /k/. All other allophone forms mentioned above are possible for all the stops.

We digitized the data, as previously described, and to use cycle-by cycle parameters in order to ascertain precisely the beginning and end points of each allophone as well as the boundaries of all acoustically distinct subphonetic segments, such as transition cycles. Exact segmentation is crucial for computing the best possible acoustic characteristics of the segments themselves. This work required the *parameters of each cycle* of the waveform to be examined, individually, by hand. We performed this segmentation chiefly on the basis of the cycle-by-cycle time-domain parameters, along with the aid of LIP plots, waveforms, and listening to the audio tapes. The times of the individual acoustic segment boundaries within each allophone were noted on the initial up-crossing of the first cycle of each acoustic segment. For fricatives, these acoustic segments included voicing regions, unvoiced frication, voiced frication, and transition regions (both individual transition cycles as well as more gradual extended transitions) in and out of frication regions. For stop consonants, these included stop preview transitions, stop previews, voicing regions, pause cycles, release-aspiration, and transitions (both individual transition cycles as well as longer transitions) at the conclusion of release-aspiration. As previously described in Chapter II, these acoustic segments may be ascertained from abrupt changes of certain kinds occurring in one or more of the cycle parameters. The cycle parameters we examined for this detailed hand segmentation, consisted of cycle-frequency, cycle-amplitude, and microstructure. After this work was completed, all the acoustic segment boundaries had to be entered into files. These files were then used to direct a battery of statistical tests to be performed on each of the acoustic segments specified.

The results could be used to compare the acoustic characteristics of like segments for all the allophones. And such quantitative results may then address a number of issues. such as:

1) What acoustic differences exist generally between /p/, /t/, and /k/, or between the low energy fricatives /v/ and /ð/?

2) How may a retroflexed /t/ be distinguished from a non-retroflexed /t/?

3) Which has the greater effect on /s/ duration, nasalization or rounding?

Another important question we may ask is whether specific allophone effects are consistent from one phone to another in the same phone class. For example, does a nasalized /t/ as compared to a non-nasalized /t/, bear a similar relationship to a nasalized /p/ compared with a non-nasalized /p/? In short, are there regular acoustic transformations which characterize the nature of the different allophone classes themselves? With respect to vowels, similar kinds of questions and methodology were adopted by Gerstman [G1], who found that classification of a single speaker's vowels in two-formant space, could be accomplished given only the first two formants of 2 or 3 known referent vowels, and the first two formants of each of the others.

We have performed pair-wise phone recognition tests, based on the allophone effects derived from single instances of allophones from the combined set of /b/, /p/, /g/, and /d/ by the same speaker. These recognition tests compared 6 measures of the release-aspiration segments of the stop consonants. For the comparison of fricatives, the same 6 measures were used to characterize frication acoustic segments. Except for a duration measure of the acoustic segment itself, the other measures were averages of the individual cycle parameters observed during the course of the whole acoustic segment. The 6 measures utilized were 1) cycle-frequency, 2)cycle-frequency dispersion, 3) duration, 4) total variation, 5) microstructure, and 6) absolute amplitude. The cycle-frequency dispersion measure has not been previously mentioned; it is simply the standard deviation of all the cycle-frequency values observed within a given acoustic segment. The value of this measure is, of course, higher for a voiced frication segment, for example, of /z/ as compared to the unvoiced frication segment of its counterpart /s/. In the former case, there are voicing cycles with much lower cycle-frequency values mixed in with high cycle-frequency values, and in the latter case only high cycle-frequency values are observed.

For these tests, rare and occasionally occurring allophones were excluded because these allophones were generally unfamiliar to the speakers themselves, and were therefore prone to articulation errors. In addition, since many of them applied only to a single phone or to a single voiced-unvoiced phone pair, they could not be used for general phone comparisons. In addition, release-aspiration segments containing a total of 3 or fewer cycles were excluded from consideration due to the unreliability of such small sample size in standard deviation computations and so forth.

For the stop consonants, the referent allophone against which all the other allophones were compared, was the aspirated form of each phone. For the fricatives, the referent used for each phone, was the the unmodified form. For each speaker then, we computed, for each of the 6 measures of each allophone, its ratio with the corresponding measure of the appropriate referent allophone.

Given the hypothesis that, for example, the release-aspiration portions of the stop consonants are inherently acoustically characterized by such measures as we have chosen, and that regular acoustic transformations of allophones exist, and may be expressed relative to the individual phone referents, we performed the following experiment. For a given speaker, we have 1 instance each, of the aspirated allophones of /g/, /k/, /b/, /p/, /d/, and /t/. These are our referents. As previously described, we compute for each measure of each of the other allophones, the ratio $R_{m,p}(i)$, of its value, $V_{m,p}(i)$, to the corresponding value of its referent allophone, $V_{o,p}(i)$.

1) $R_{m,p}(i) = V_{m,p}(i)/V_{o,p}(i)$

where $V_{m,p}(i)$ = value of parameter i for allophone m of phoneme P

and allophone m = o represents the referent value.

If the hypothesis described above is correct, we hope to discriminate with distance measures, between 2 phones, regardless of which allophone of these two is being tested, solely on the basis of knowing 1) the aspirated referents for each of the 2 test stop consonants, and 2) the average relationship of each allophone type, as derived from the combined allophone ratios of the training set of the other 4 phonemes, not being tested.

Therefore:

2) $A_m(i) = (\Sigma_{p=1,q} R_{m,p}(i)/q)$

where q = number of phonemes in training set.

3) $\sigma_m(i) = (\Sigma_{p=i,q}(R_{m,p}(i) - A_m(i))^2 / (q - 1))^{1/2}$

We may then predict a prototype value of parameter i for a given allophone type, for any phoneme for which we have a referent value of parameter i.

4) $P_m(i) = V_{o,q+1} A_m(i)$

where phoneme p = q + 1 represents the phoneme for which the predicted prototype is to be computed.

Finally we take a modified Euclidean distance measure between each of the predicted allophone parameter values and the test phoneme sample parameter values

5) $T_m = (\Sigma_{i=1,6}((P_m(i) - S_m(i)) / \sigma_m(i))^2)^{1/2}$

where $P_m(i)$ = predicted prototype value of parameter i

$S_m(i)$ = test phoneme sample value of parameter i

$\sigma_m(i)$ = standard deviation of parameter i

T_m = distance measure of allophone m of selected phoneme.

With this measure, chance predicts 50% correct choices and 50% incorrect choices, in choosing one of the two test phonemes. Note that this test paradigm is very strenuous for several reasons. For each speaker, only one example of each allophone exists in the data base. Of necessity, the sample space from which to derive allophone effect statistics is very limited. Therefore, in discriminating between any 2 stop consonants, only 4 samples (one from each of the other stop consonants), at best, are available for providing the derivation or training for each of the different allophone characteristics.

Clearly, any errors in this training set may assume a large effect. In the original allophone classification of the data, the assumption was made that if a phone occurred in a given environment, it was necessarily coarticulated with that environment, and therefore was labeled as such. However it is not necessary that a phone be coarticulated with its environment; it may, and in fact often does, remain unmodified by its environment. Nevertheless its allophone label is the same as though it were coarticulated. For example, in the utterance "sip more", the /p/ is labeled in the data as being a nasalized allophone of /p/, whether it is acoustically nasalized or whether it is acoustically the same as the unmodified allophone of /p/ in the utterance "sip it". Therefore, it is very likely the case, that unmodified allophones are averaged in with modified allophones in the training sets for the various kinds of modified allophones, and contribute a source of error.

The results of this experiment, however, are surprisingly good. We obtained scores for the tests of pair-wise phone discrimination, both for phones with same and different places of articulation. Of particular interest are the results for discrimination of phones with different places of articulation, such as between /p/ and /t/, for example. Therefore in the tests for all allophones of 1) /g/ and /k/ vs the allophones of /b/, /p/, /d/, and /t/, 2) /b/ and /p/ vs /g/, /k/, /d/, and /t/, and 3) /d/ and /t/ vs /g/, /k/, /b/, and /p/, the following scores were obtained:

Different Place of Articulation

Male 1 - 75%

Male 2 - 86%

Female 1 - 71%

These tests assume that no knowledge is available concerning the acoustic environment in which an unknown phone occurs. This is, in fact, hardly ever the case in speech segmentation and/or labeling processes. Recall that in these tests, an unknown allophone, belonging to one of the two phonemes tested, is being compared against a predicted set of parameter values for each allophone of both test phonemes, and that the specific allophone, characterized by these predicted values, with the shortest distance measure from the unknown allophone determines the test phoneme chosen by each test. With speech processing programs, it is very often the case that information is available concerning the nature of the actual context of an unknown stop consonant, for example. In all likelihood, given such information, the probability of confusing similar stop consonants would be diminished. This kind of redundancy could prove to be quite useful for accurate labeling. For example, if in analyzing an unknown stop consonant, the distance measures for a nasalized /p/ and a rounded /k/ were close, one could look for contextual evidence of a nasal to help resolve labeling confusion. This kind of contextual test is generally not difficult, since it is not necessary that the contextual phones be actually identified, but only their general phone classes. However, it is also likely that for optimal labeling, accurate statistics are required on individual speaker coarticulation phenomena.

Next, we examine the less critical discrimination of stop consonants with the same place of articulation; specifically, /g/ vs /k/, /b/ vs /p/, and /d/ vs /t/. Although all the voiced stops were preceded by voicing, we did not use this important cue for discrimination, but rather compared only the release-aspiration segments.

The scores obtained for our three speakers follow:

Same Place of Articulation

Male 1 - 63%

Male 2 - 69%

Female 1 - 44%

These much lower discrimination scores indicate the high degree of acoustic similarity between the release-aspirations of voiced-unvoiced phone pairs. In many cases, a voiced stop goes de-voiced at or very near the start of its release-aspiration. After that point, the release-aspiration of the voiced stop is very similar to that of the unvoiced phone with the same place of articulation. However, it is also true, that certain consistent differences generally do prevail between voiced phones and their unvoiced counterparts, the shorter duration of the voiced phones, for example. In addition, we note that the discrimination scores for the female speaker were lower than those of the two male speakers. This may be due to the higher ambient noise level in the recordings of her voice.

We performed the same pair-wise phone recognition tests among the 8 fricatives previously enumerated. For the fricatives then, the training set consisted of all the allophone data of 6 fricatives The scores follow for the pair-wise discrimination of fricatives with different places of articulation; that is, 1) /s/ and /z/ vs /ʃ/, /ʒ/, /θ/, /ð/, /f/, and /v/, 2) / / and /ʒ/ vs /s/, /z/, /θ/, /ð/, /f/, and /v/. 3)/θ/ and /ð/ vs /s/, /z/, /ʃ/, /ʒ/, /f/, and /v/, and 4) /f/ and /v/ vs /s/, /z/, /ʃ/, /ʒ/, /θ/, and /ð/.

Different Place of Articulation

Male 1 - 95%

Male 2 - 91%

Female 1 - 83%

As for the stop consonants, contextual information was not used for these phoneme discrimination tests. In actual practice, we expect that use of such generally available information would improve phoneme discrimination or labeling even further. Discrimination scores for fricatives with the same place of articulation; that is, /s/ vs /z/, /ʃ/ vs /ʒ/, /θ/, vs / /, and /f/ vs /v/, as expected, are somewhat lower.

Same Place of Articulation

Male 1 - 83%

Male 2 - 75%

Female 1 - 75%

We note that, on the whole, the discrimination scores for all the speakers, both for phonemes with the same and those with different places of articulation, are somewhat higher for the fricatives than for the stop consonants. This may be due in part to increased stability with the larger set of training data for the fricatives, with 6 phonemes rather than 4. Another contributing factor for the difference may be the fact that, on the average, the duration of frication is substantially longer than that for the release-aspiration of stops. And statistics gathered over longer duration segments may be more stable. Specifically, in examining the release-aspiration segments of all the stop consonant allophones, we find that 10% of them last for less than 5 msec in duration, and 8% of them last for less than 3.3 msec (usual wide band spectrogram resolution) in duration. In normal continuous speech, these percentages are probably higher yet. On the other hand, only 1% of the frication segments were less than 10 msec in duration.

This table summarizes all the results just presented.

Summary of Allophone Tests

	Male 1	Male 2	Female 1
Stops:			
Diff. Place	75% (60/80)	86% (60/70)	71% (49/69)
Same Place	63% (12/19)	69% (11/16)	44% (7/16)
Fricatives:			
Diff. Place	95% (137/144)	91% (118/129)	83% (100/120)
Same Place	83% (20/24)	75% (15/20)	75% (15/20)

In Appendix C there are matrices presented which show for each speaker the specific pair-wise phone comparison test results, with the results combined for different and same place of articulation.

The following tables illustrate the significance of using allophone information for identifying phonemes accurately. Taking each parameter in turn, we show for each of the three speakers, the relationships of the phonemes characterized by these parameters. First we give the range of values assumed by each of the phonemes. Clearly there is a great deal of overlap among these phoneme ranges. However, if we compute the average ratio of each allophone type to its phone referent, we find that the different allophone types each adhere, rather consistently, to specific regions within these ranges. Therefore the search space for a given allophone is significantly narrowed. So next we present the allophone ratios for all the phones of all the speakers combined. Note that the separation of these allophone ratios for the fricatives is much smaller than for the stop consonants. Separate analyses of averages and standard deviations for all the allophone types, for each speaker, are included in the Appendix. Note also that although, for a given parameter, two allophone types may occupy similar positions, for another parameter, these allophone types may occupy distinctly different positions. It is the total pattern of the different parameter relationships which characterize any given allophone. In these tables, the following abbreviations are used for the different allophone types:

- unmodified

A palatalized

C minimally released

H aspirated

N nasalized

R retroflexed

T dentalized

W rounded

STOP CONSONANTS

MALE #1

	CYCLE-FREQ	CF DISP	DURATION	TV(NORMALIZED)	MICROSTRUCTURE	ABSAMP
PHONE	P #1(Hz)	P #2	P #3(msec)	P #4	P #5	P #6
	min-max	min-max	min-max	min-max	min-max	min-max
/g/	1[illegible]24-2638	622-1702	13.2-82.9	317-890	73-315	33-97
/k/	1214-4111	484-1700	23.2-74.3	536-1233	95-241	44-136
/b/	996-2499	103-1119	3.0-6.0	553-774	69-170	43-67
/p/	1663-2485	93-1139	2.9-33.4	618-740	66-186	55-81
/d/	1501-3778	648-2148	1.3-35.1	569-1193	116-264	13-119
/t/	1636-4491	714-3352	10-4-101.2	676-1453	130.494	22-132

Male #2

/g/	1059-2443	121-962	11.4-50.0	342-830	22-123	55-104
/k/	1251-2610	148-469	3.8-154.6	414-834	31-78	32-214
/b/	1562-2180	478-673	1.9-29.7	628-756	97-135	31-62
/p/	1574-2383	547-894	3.4-110.2	609-795	82-120	29-110
/d/	2144-2874	754-1648	1.9-60.0	721-952	108-261	23-70
/t/	1649-3218	546-1544	10.4-122.4	698-1088	82-310	26-87

FEMALE #1

/g/	607-2843	184-1161	11.9-24.3	489-571	41-443	59-162
/k/	1068-3131	228-550	26.2-90.2	359-568	43-332	24-104
/b/	982-1867	261-1154	3.2-20.2	345-730	65-207	18-58
/p/	1028-1939	268-1278	10.4-32.8	413-708	64-147	24-45
/d/	1410-3739	535-1597	1.0-36.9	493-579	58-245	41-103
/t/	1189-2708	744-1167	6.9-88.7	428-794	43-159	34-138

AVERAGE ALLOPHONE RATIO PERCENTAGES

P #1		P #2		P #3		P #4		P #5		P #6	
N	66%	N	74%	C	27%	N	80%	R	60%	T	52%
R	74	R	84	N	46	R	85	N	68	C	81
W	84	C	90	W	51	C	91	C	92	A	90
C	96	H	100	A	53	W	93	H	100	N	91
H	100	W	123	-	53	H	100	W	107	H	100
-	108	-	156	T	59	-	106	-	117	R	108
T	110	T	163	R	76	T	118	T	162	-	110
A	219	A	186	H	100	A	144	A	187	W	122

FRICATIVES

Male #1

PHONE	CYCLE-FREQ P #1(Hz) min-max	CF DISP P #2 min-max	DURATION P #3(msec) min-max	TV(NORMALIZED) P #4 min-max	MICROSTRUCTURE P #5 min-max	ABSAMP P #6 min-max
/f/	3293-5663	2177-3206	106.3-184.7	1186-1494	315-360	32-38
/v/	1258-3772	693-2564	28.7-62.2	672-1177	235-323	42-63
/θ/	2667-3286	1704-2017	135.6-186.4	1079-1249	324-408	27-32
/ð/	1517-2688	1053-2069	13.1-92.2	674-920	201-348	43-52
/ʃ/	3678-3849	1559-1627	190.2-237.1	1092-1137	169-194	168-217
/ʒ/	3112-3676	1183-1865	92.2-140.8	99[illegible]-1090	147-193	153-256
/s/	4561-6328	1121-1529	151.2-204.3	1303-1521	212-265	146-284
/z/	4738-4976	1253-1472	80.9-158.9	1240-1333	215-234	114-191

Male #2

PHONE	CYCLE-FREQ P #1(Hz) min-max	CF DISP P #2 min-max	DURATION P #3(msec) min-max	TV(NORMALIZED) P #4 min-max	MICROSTRUCTURE P #5 min-max	ABSAMP P #6 min-max
/f/	1871-2646	927-1129	126.7-211.5	775-913	175-234	31-42
/v/	1964-2376	987-1325	41.7-75.1	769-914	186-256	28-44
/θ/	1607-2520	817-1298	72.6-182.0	826-1011	92-265	26-34
/ð/	1637-2112	565-1200	16.5-28.7	656-833	102-204	28-76
/ʃ/	2678-3158	864-1097	196.8-274.6	857-976	105-132	218-285
/ʒ/	2366-3167	723-1057	127.3-174.7	796-972	104-130	112-157
/s/	3735-4771	725-1186	199.5-294.3	1221-1302	174-223	74-110
/z/	4175-4674	689-1355	116.0-183.4	1170-1259	183-221	42-86

Female #1

PHONE	CYCLE-FREQ P #1(Hz) min-max	CF DISP P #2 min-max	DURATION P #3(msec) min-max	TV(NORMALIZED) P #4 min-max	MICROSTRUCTURE P #5 min-max	ABSAMP P #6 min-max
/f/	1280-2562	789-1039	34.5-160.5	632-909	182-214	24-32
/v/	556-1118	260-472	9.4-20.0	329-373	38-71	42-77
/θ/	980-1390	681-847	42.7-46.4	550-561	146-235	20-59
/ð/	647-1371	651-815	40.5-72.9	364-470	89-134	38-268
/ʃ/	2820-3744	576-1526	223.1-262.7	893-1097	100-157	128-159
/ʒ/	1512-2905	1017-1135	70.7-118.4	528-926	124-160	82-104
/s/	2690-4564	1248-1984	90.4-235.2	973-1264	218-361	27-132
/z/	1740-4353	1406-2659	31.7-113.1	693-2178	174-304	37-83

AVERAGE ALLOPHONE RATIO PERCENTAGES

P #1		P #2		P #3		P #4		P #5		P #6	
T	99%	N	95%	T	74%	W	97%	R	90%	T	73%
W	100	-	100	R	87	T	100	-	100	W	98
R	100	T	100	N	88	-	100	N	101	-	100
-	100	W	107	W	91	N	101	W	102	R	107
N	105	R	110	-	100	R	110	T	117	N	133

MUSIC - VIOLINS

Of course, since time-domain analysis is generalizable to any waveform, we may also examine non-speech signals, with the hope of addressing questions of characterization in general, as well as specific issues re: transient phenomena and temporal measures. For example, in conjunction with Schumacher [B3], we have briefly examined some violin and cello music recorded under recital conditions and digitized at 10 kHz. Generally the waveforms of these instruments resemble the vowel portions of human speech. "Pitch periods" are readily apparent throughout. However the time-domain analysis reveals that these pitch periods are not perfectly regular. The individual cycles within a given pitch period may fluctuate slightly from those in successive pitch periods. Rapid fluctuations of this type have been referred to as "jitter" in perception experiments [B9]. This jitter is evident in the accompanying LIP plot (Fig. 45) of the open G string (G3), of the violin, which at about 1.53 sec is followed by the first finger position (A3), played on the G string, with vibrato.

Note first that the clear horizontal lines are not perfectly straight and regular. The jitter of successive pitch periods is reflected in the log inverse period or cycle-frequency measure displayed. Perception experiments by Pollack [P2] indicate that such jitter may be perceptible to human listeners. In fact, it may well be that these irregularities are a distinguishing feature of instrumental music. For example, we find this difference between true violin music and synthesized violin music which is built up as a simple composite of sine waves or other regular waveforms. In the waveform displayed above the LIP plot, vertical lines have been automatically drawn pitch synchronously. Following each line is a number measuring the duration of the previous period. This measure is in units of centiseconds; therefore a 50 centisec duration corresponds to a frequency of 200 Hz.

A longer term fluctuation is that caused by the vibrato. The rapid changes of finger pressure create both amplitude and cycle-frequency modulations. The A3 note commencing at about 1.53 sec is played with vibrato. Although not much of this note is shown here, the oscillations of cycle-frequencies can be seen. The waveform amplitude modulations are more easily seen in the expanded waveform shown in Fig. 46. The full expanded waveform shown here lasts 2 seconds, and should be read from left to right, and from top to bottom. For any given line, the duration of time between vertical lines is 40 centiseconds (.04 seconds). The note A3 commences at about 206 centiseconds. Here we see about 2 1/2 periods of amplitude modulation in the waveform envelope for A3. The musician playing vibrato here typically incorporated 5 or 6 periods of amplitude modulation per second. For comparison, observe the lack of such periods during the playing of the open G string. The only major change in amplitude seen here is a gradual increase, which reflects increased bowing pressure. The LIP plot may be directly related in time to the expanded waveform if one equates the 50 centisecond origin of the LIP plot and converts centiseconds to seconds, or vice versa. Therefore, the 1.0 second marker on the LIP plot corresponds to the 150 centisecond marker on the expanded waveform.

Next, note the apparently sharp discontinuity in LIP pattern occurring at about .67 sec. This change does not represent a change of note. Examination of the waveform above the LIP plot shows that both before and after this discontinuity, the pitch period durations are essentially the same. We know that the friction of bowing excites the inherent resonances of the violin. The discontinuity seen here is probably a threshold response to the gradual change of bowing pressure, which elicits a shift in the relative amplitudes of the cycle-frequency components within the pitch period.

Finally, note in the LIP plot, the large irregularities occurring at major transition regions. These transition regions reflect 1) the beginning of bowing at about .23 sec and 2) the change in note and bowing direction at about 1.53 sec. It is known that such transitional regions provide important cues for musical instrument identification.

From an analysis point of view, it may be possible that perception experiments based on digital manipulation of such transition regions could elucidate information-bearing components of these. More generally, time-domain analysis may prove helpful in providing better descriptions of individual musical instruments as well as clarification of certain musical phenomena; e.g. jitter. Alternatively, the incorporation of some of these phenomena may allow a better synthesis of certain musical instruments than is presently available.

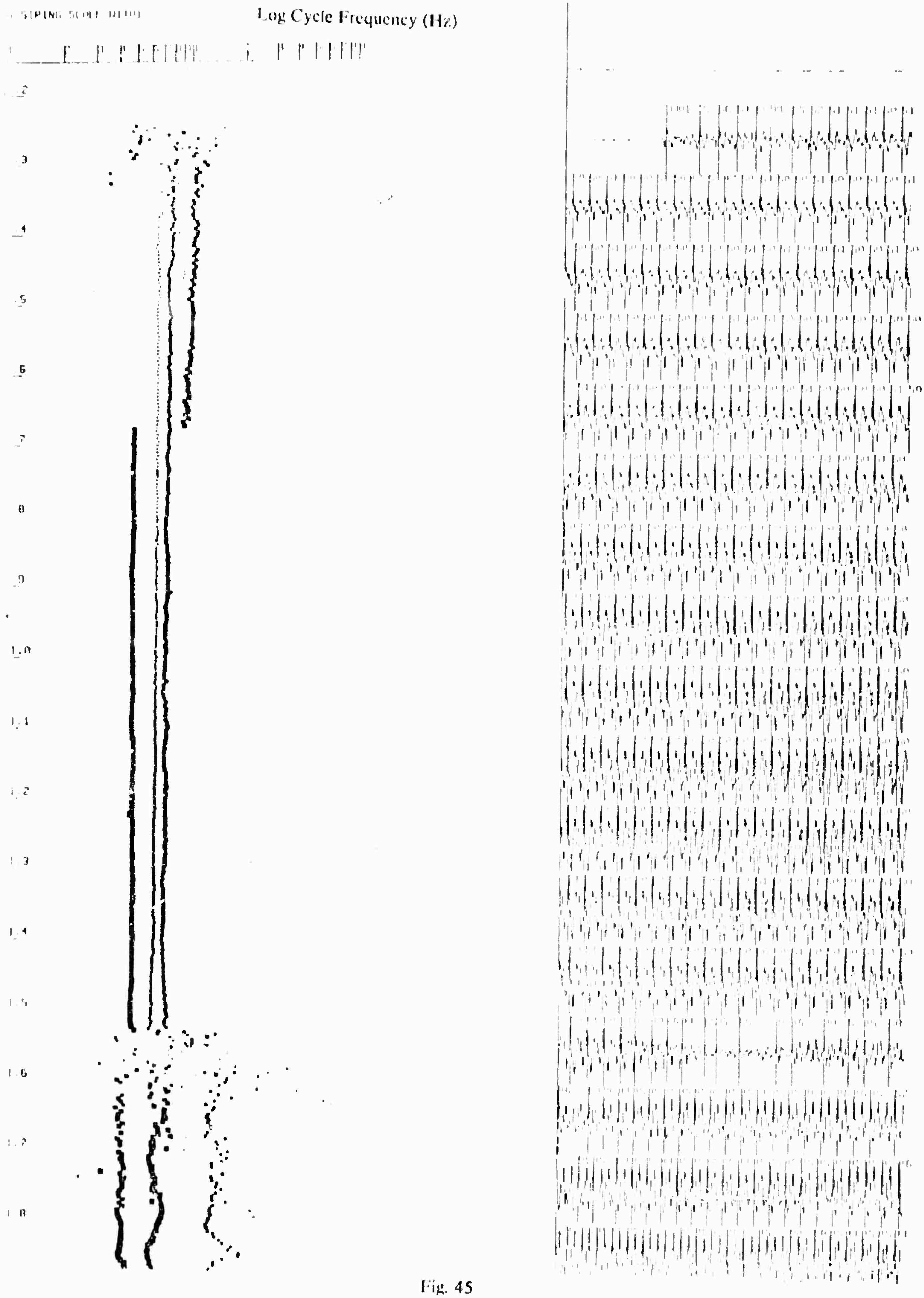


Fig. 45

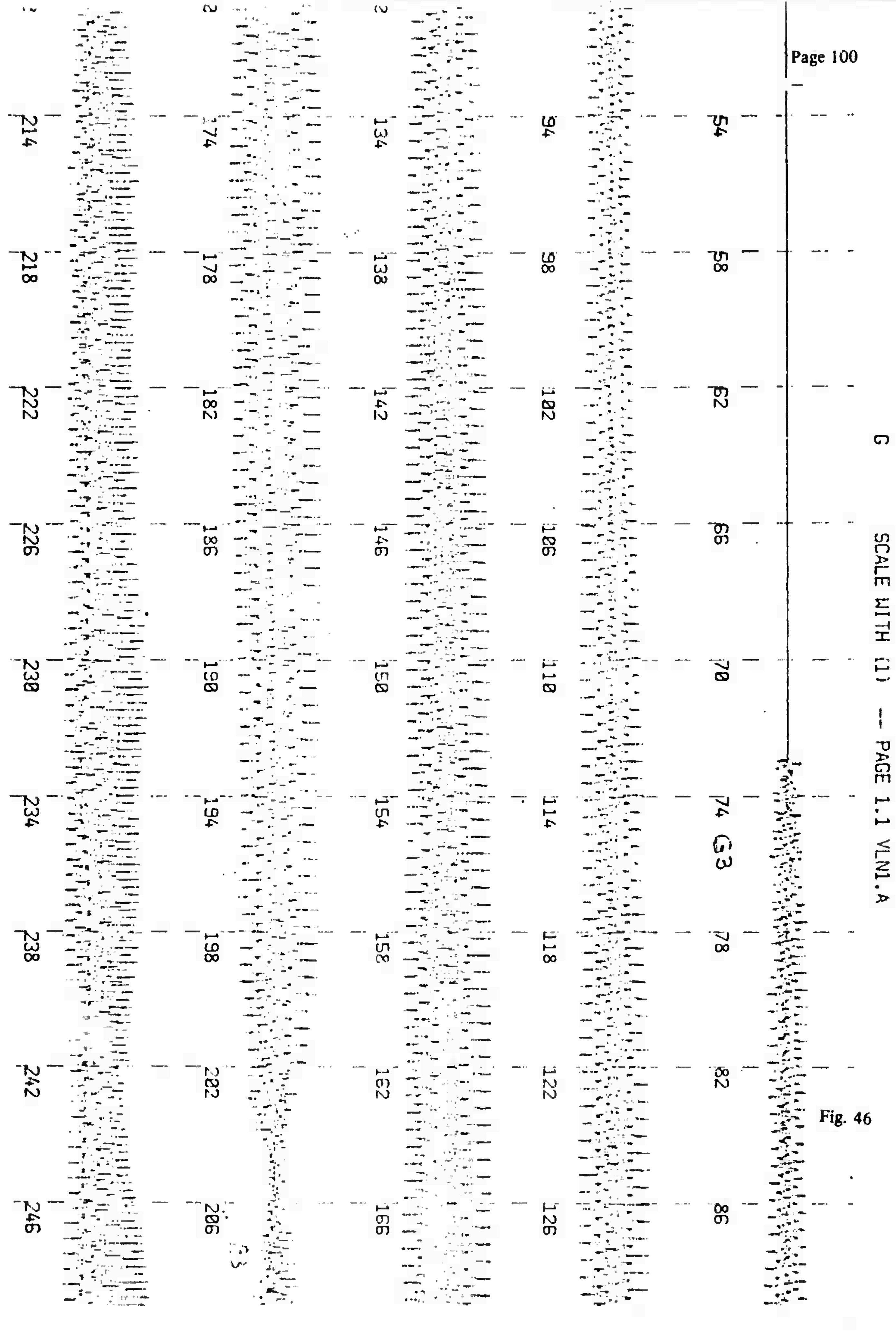

G SCALE WITH (1) -- PAGE 1.1 VLN1.A
54
58
62
66
70
74 G3
78
82
86
94
98
102
106
110
114
118
122
126
134
138
142
146
150
154
158
162
166
174
178
182
186
190
194
198
202
206
214
218
222
226
230
234
238
242
246

Fig. 46

ANIMAL VOCALIZATIONS - BOU-BOU SHRIKES

The bou-bou shrike (*Laniarius aethiopicus*) which is also known as the bell shrike or bellbird, is native to East Africa. Its song, largely devoid of harmonic content, is almost sinusoidal acoustically and sounds quite melodic to the ear. Surprisingly, not all the individual notes in the song are produced by a single individual bird, but rather by two separate birds whose vocalizations are highly synchronized. This fact is apparent however to an observer standing between the two birds and hearing the notes emanating from separate directions. This kind of song, usually begun by one bird and finished by the other, is called "antiphonal song", a form of the more general duetting phenomena. The separate vocal contributions of the two birds may overlap in time, completely, partially, or not at all. The repetoire of precise song patterns sung by a pair of birds may be quite large, with some patterns identical to those of other local pairs of the same species, and still other patterns unique to the specific bird pair itself. Generally this antiphonal song is characteristic of a male-female pair mated for life, with a relatively permanent year-round territory amid dense tropical foliage. It is thought [T2] that antiphonal singing operates as a mechanism for confirming and maintaining pair contacts where visual cues are practically absent. Since males and females have identical physical form and color characteristics, their song conveys the song and individual-specific information necessary for identification.

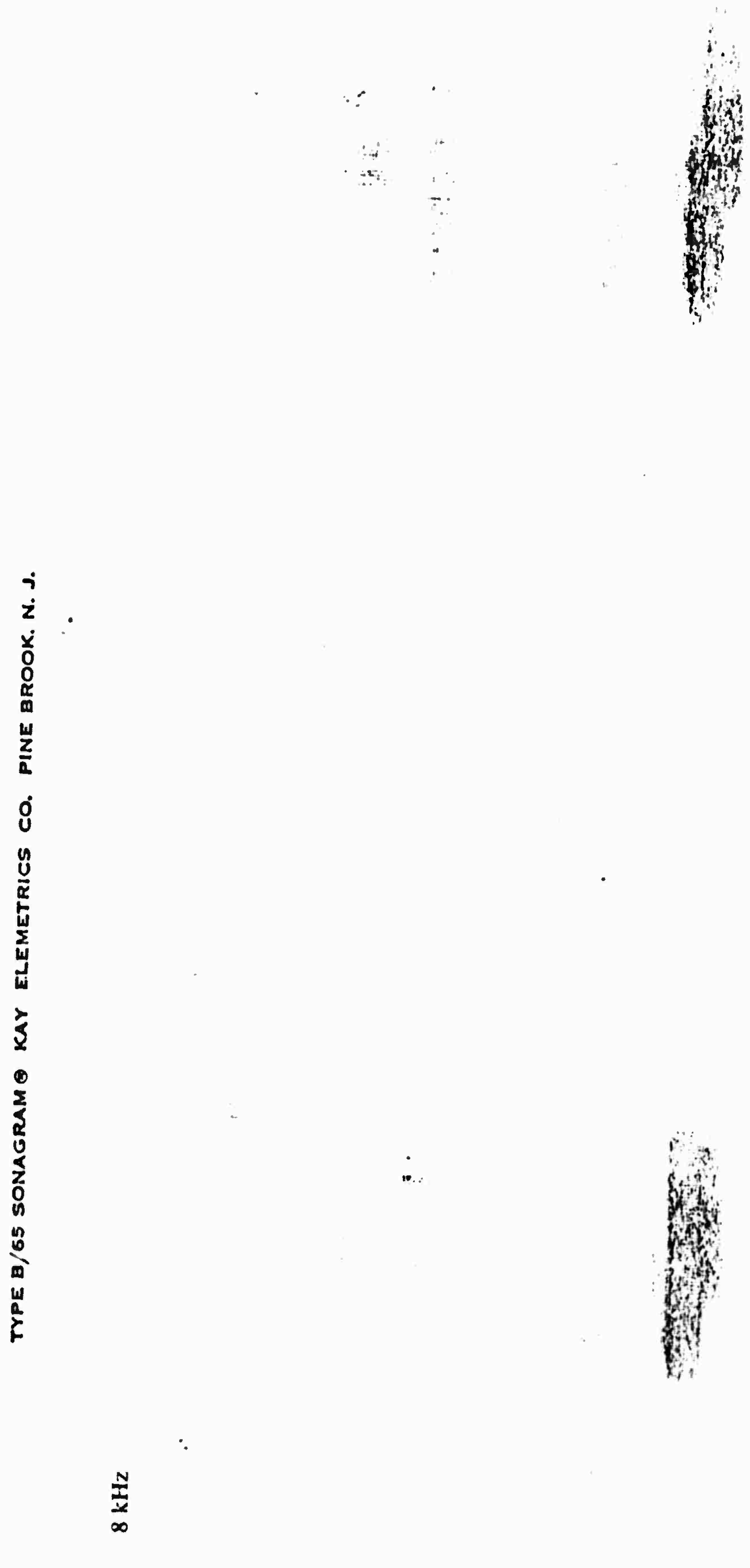
Thorpe [T1] noted that for *Laniarius erythrogaster*, a "reaction time" could be defined as the period of time between when the first bird, A, started singing, and when the second bird, B, commenced singing. He found that this reaction time was quite consistent, even though the duration of each of the bird's vocalizations could be quite variable. He reports that for a series of 8 duets by *L.e.*, a reaction time of 144 msec with a standard deviation of 12.6 msec (= 8.75% of reaction time) was observed. He also notes observations of others on 1) a series of 7 duets, *L.a.*, with a reaction time of 425 msec and a standard deviation of 4.9 msec (= 1 [illegible] of reaction time) and 2) a series of 6 duets by *Cisticola chubbi*, with a reaction time of 396 ms[illegible] standard deviation of 2.9 msec (= .73% of reaction time). Thorpe bases his measures on spectrographic data and claims accuracy within 1.5 msec.

The Cornell Library of Natural Sounds has generously made available to us recordings of various animal vocalizations. The examp' :ussed here include recordings of *L.a.* (recorded by M. P. McChesney) and *Melospiza melodia*, the song sparrow (recorded by R. C. Stein and R. S. Little, #RCS63-99j) On a series of 4 duets of *L.a.*, we have observed a reaction time of 294.4 msec with a standard deviation of 2.108 msec (= .72% of reaction time). These measures are based on our time-domain analysis, as previously described for human speech. Temporal resolution at this sampling frequency of 20 kHz is 50 microseconds.

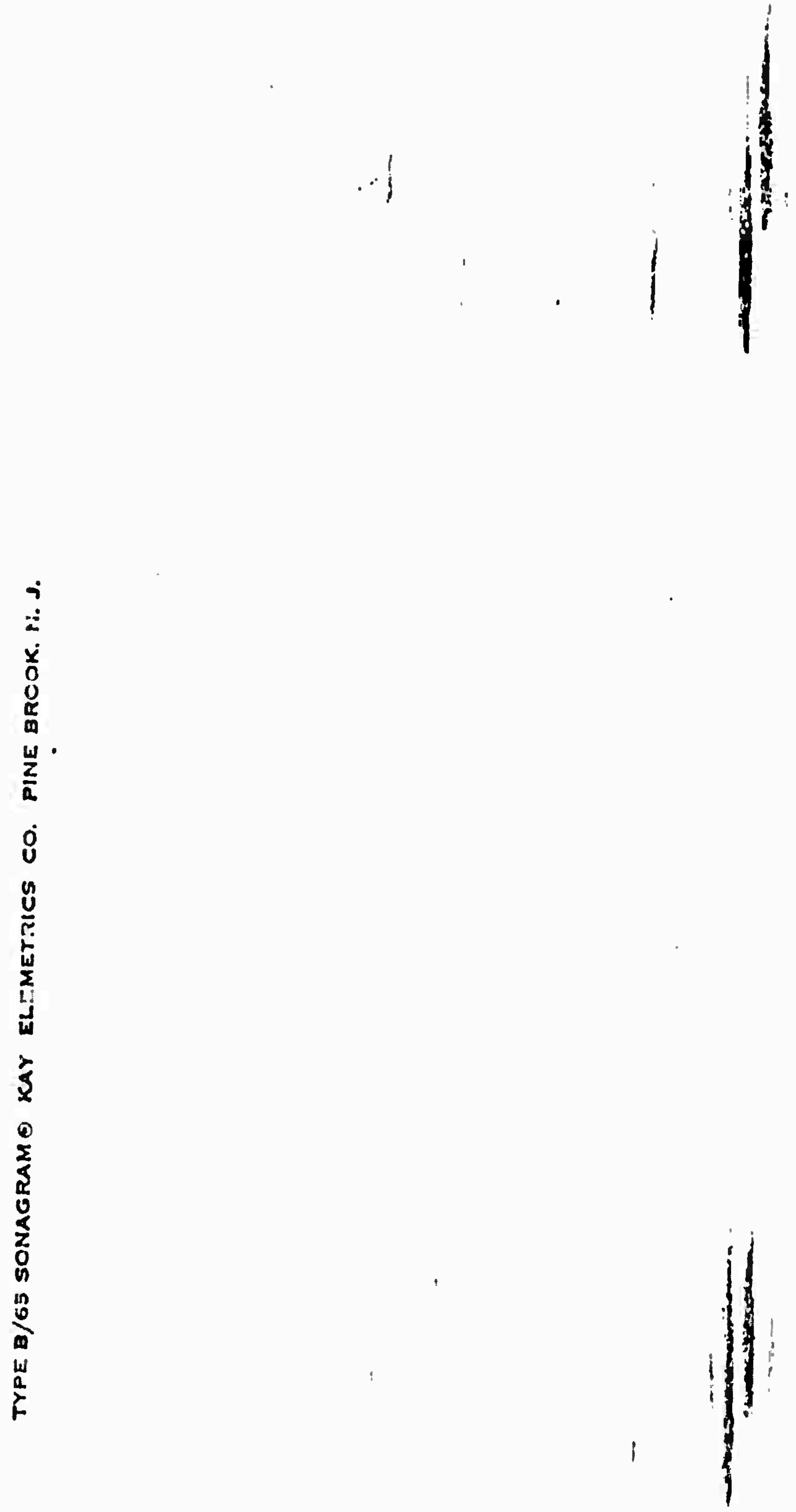
Thorpe notes that reaction times can be expected to vary as a function of 1) the differences between the individual pairs singing any given song pattern, 2) distance between the duetting birds, and 3) the current activity of the responding bird. The following illustrations show comparitively the spectographic and LIP displays for the same vocalizations. First is an example of a vocalization by *L.a.* (reaction time= 185.0 msec), portrayed by 1) a wide-band spectrogram (Fig.47), 2) a narrow-band spectrogram (Fig.48), and 3) an LIP plot (Fig. 49). Temporal resolution for each of these is 1) 3.3 msec, 2) 22.2 msec. and 3) 50 microsec, respectively. Frequency resolution on these graphical displays is 1, 300 Hz and 2) 45 Hz for the wide and narrow band spectrograms, respectively; uninterpolated cycle-frequency resolution of the LIP plot is inversely proportional to the cycle-frequency measured. At cycle-frequencies of 10 Hz, 100 Hz, and 1000 Hz, the cycle-frequency resolution is .05 Hz, .5 Hz, and 50 Hz, respectively, given the sampling rate of 20 kHz. Note that the time scale for the LIP plot is slightly reduced (3% less) than for the spectrograms, and that the y-axis of the spectrograms is linear whereas the y-axis for the LIP plot is logarithmic. The bou-bou shrike vocalization shown in the LIP plot corresponds to the first of two shown in the spectrograms.

Even in relatively noisy recordings of animal vocalizations; e.g. extraneous background bird vocalizations, time-domain analysis appears rather noise resistant and details of the primary vocalizations are usually still quite apparent. As revealed in the time-domain, bird A is singing consistently at an average cycle-frequency of 1024 Hz (cycle-frequency dispersion = 12.85 Hz which is 1.25% of the average cycle-frequency), whereas bird B is singing at an average cycle-frequency of 828 Hz (cycle-frequency dispersion = 19.92 Hz which is 2.14% of the average cycle-frequency). We have observed another bou-bou shrike which sang even more consistently, a note of 1695 Hz (cycle-frequency dispersion = 13.26 Hz which is .78% of the average cycle-frequency!). In summary, we find that both the temporal resolution and cycle-frequency resolution inherent in our time-domain techniques, are in fact required for sufficient accuracy in a detailed characterization of the bou-bou shrike vocalization. We feel that this situation is in no way unique, and that application of these techniques to other animal vocalizations may well disclose many new features, patterns, and acoustic relationships in general. In addition, this precise acoustic information (for example, the bou-bou shrike reaction time, or consistency of their song cycle-frequencies) may well reflect the capabilities of physiological receptor and production mechanisms, as they are naturally utilized.

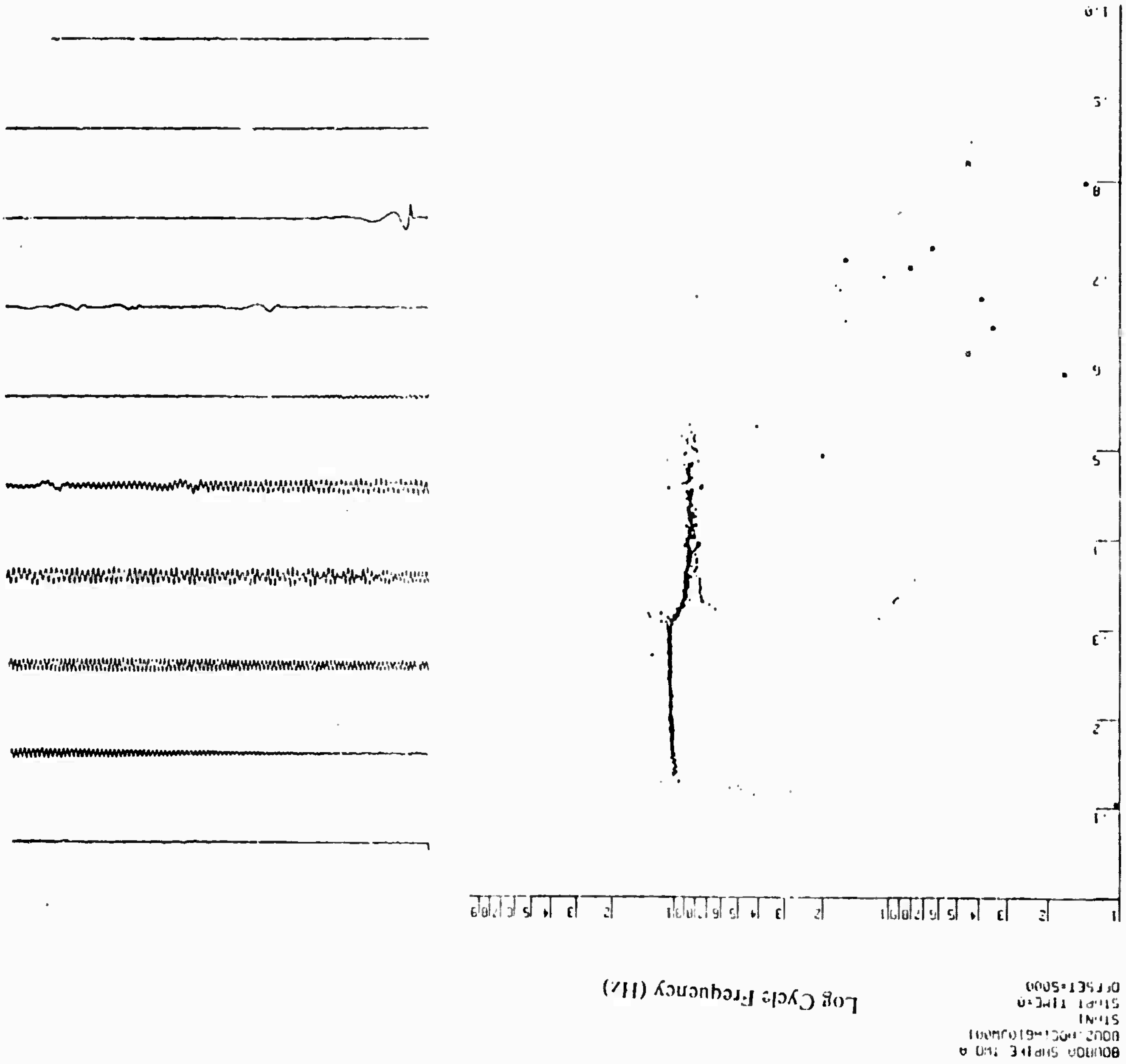
Finally we leave an exercise for the reader! The same kinds of comparative visual displays (two pages each) follow, in the same order, for the song of an individual song sparrow (Figs.50-55). This species is a familiar song bird and has a complex song characterized by both transients and steady states. Here with higher frequencies present than in the case of the bou-bou shrike, more compensation must be made for the difference between the linear and logarithmic scales . In comparing the different displays, note differences in details, many of which occur in consistent patterns, as well as differences in the gross features.



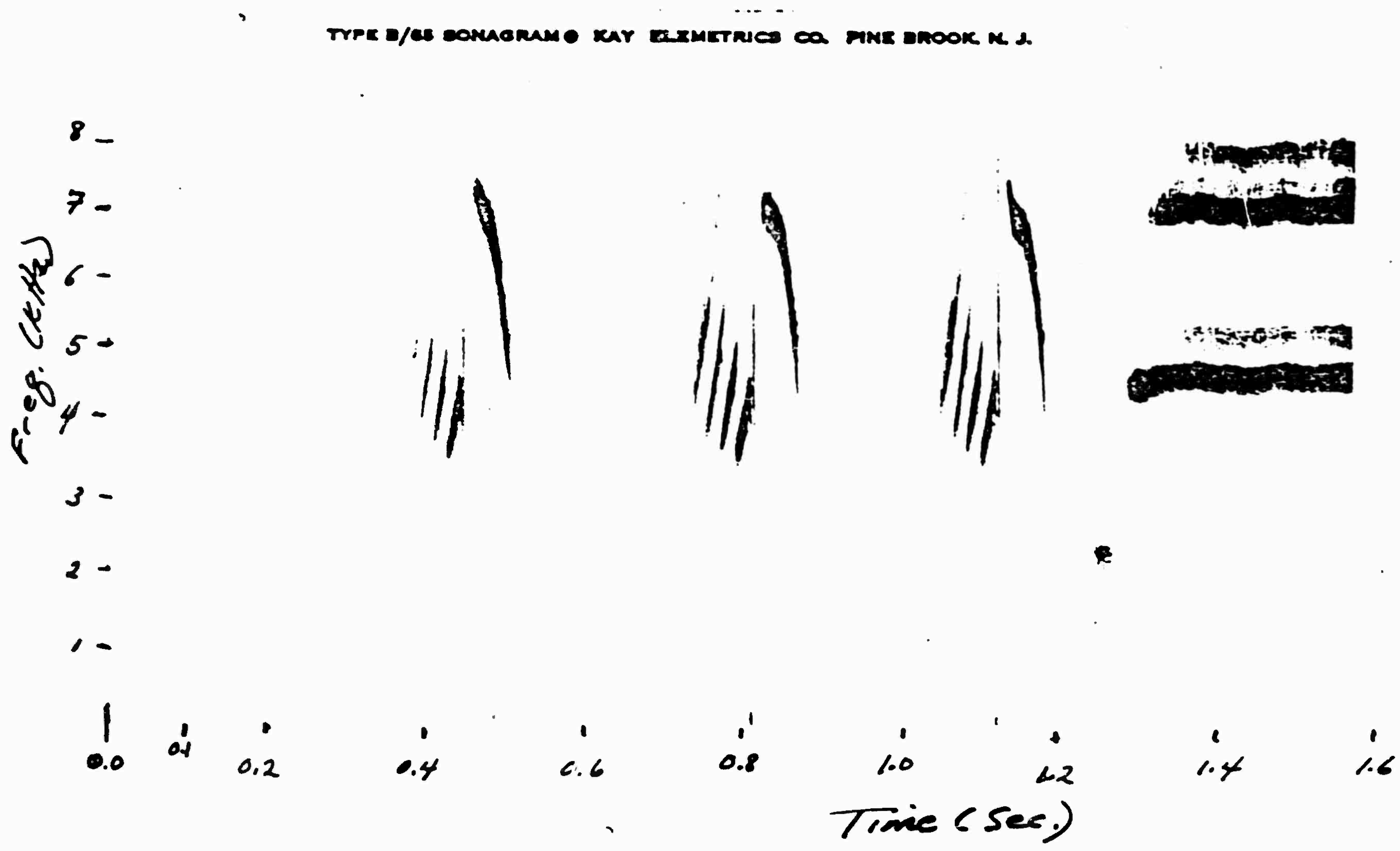
BOU-BOU SHRIKE (*Laniarius aethiopicus*)



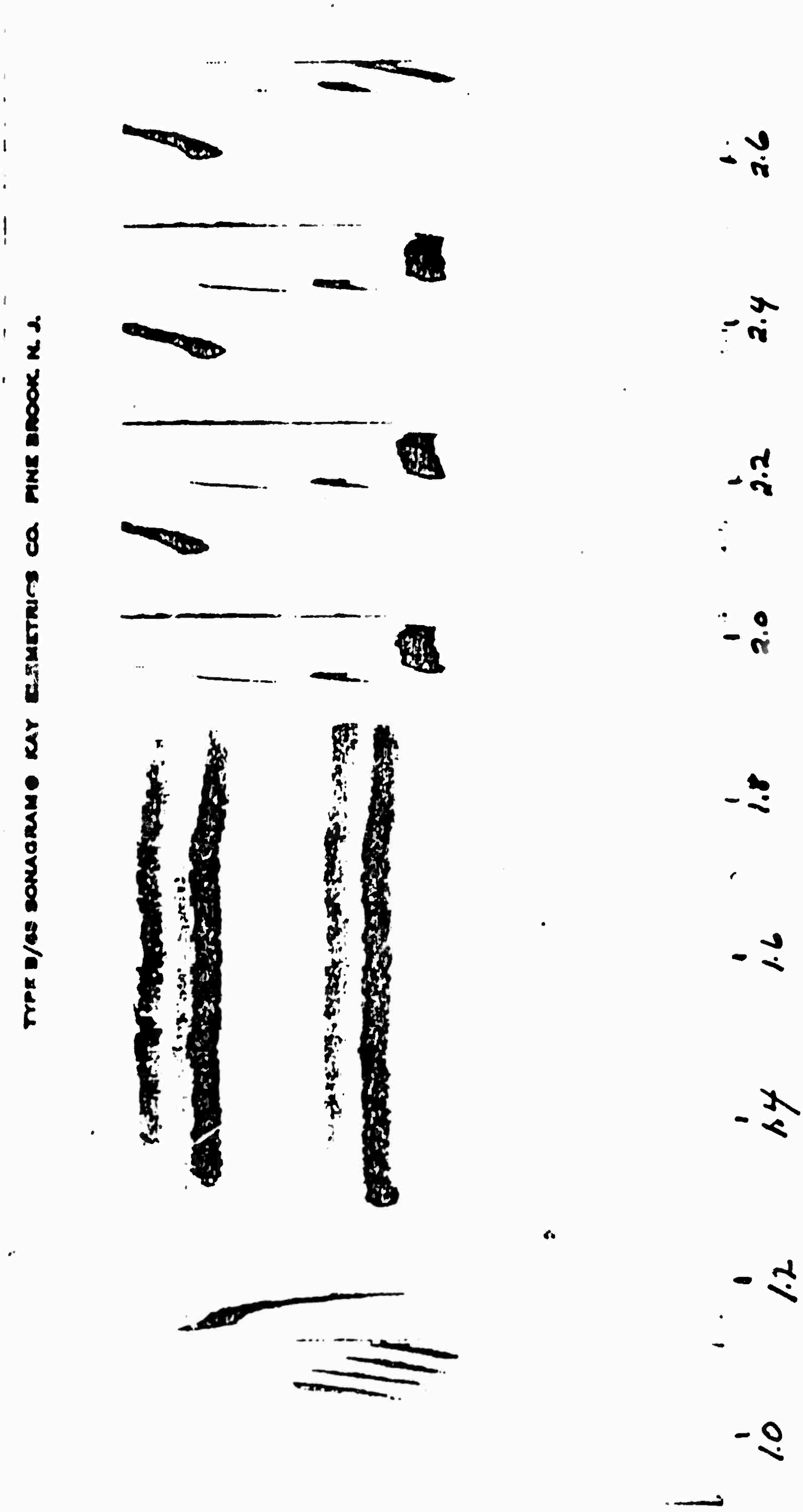
BOU-BOU SHRIKE (*Laniarius aethiopicus*)



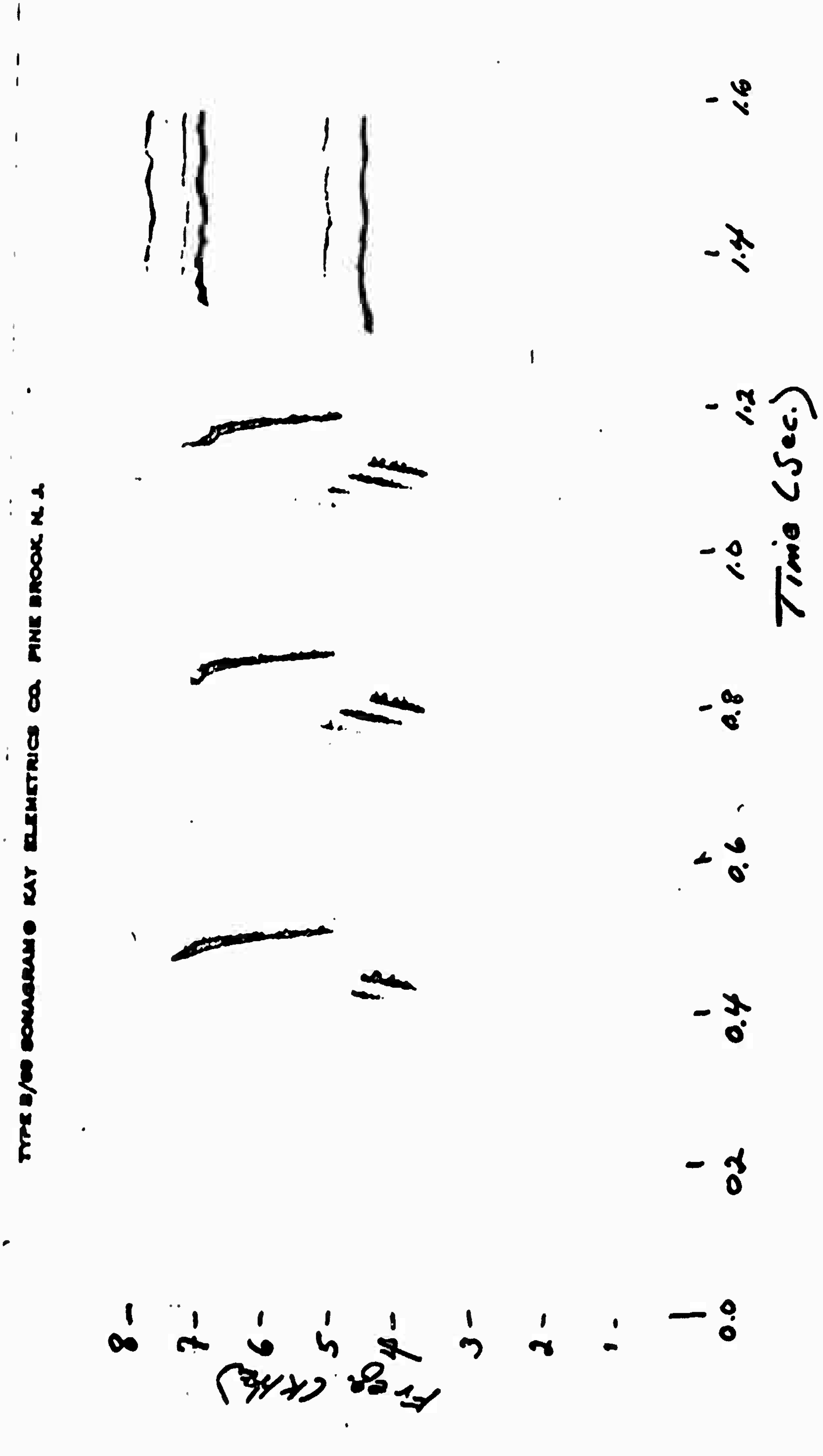
BOU-BOU SHRIKE (*Laniarius aethiopicus*)



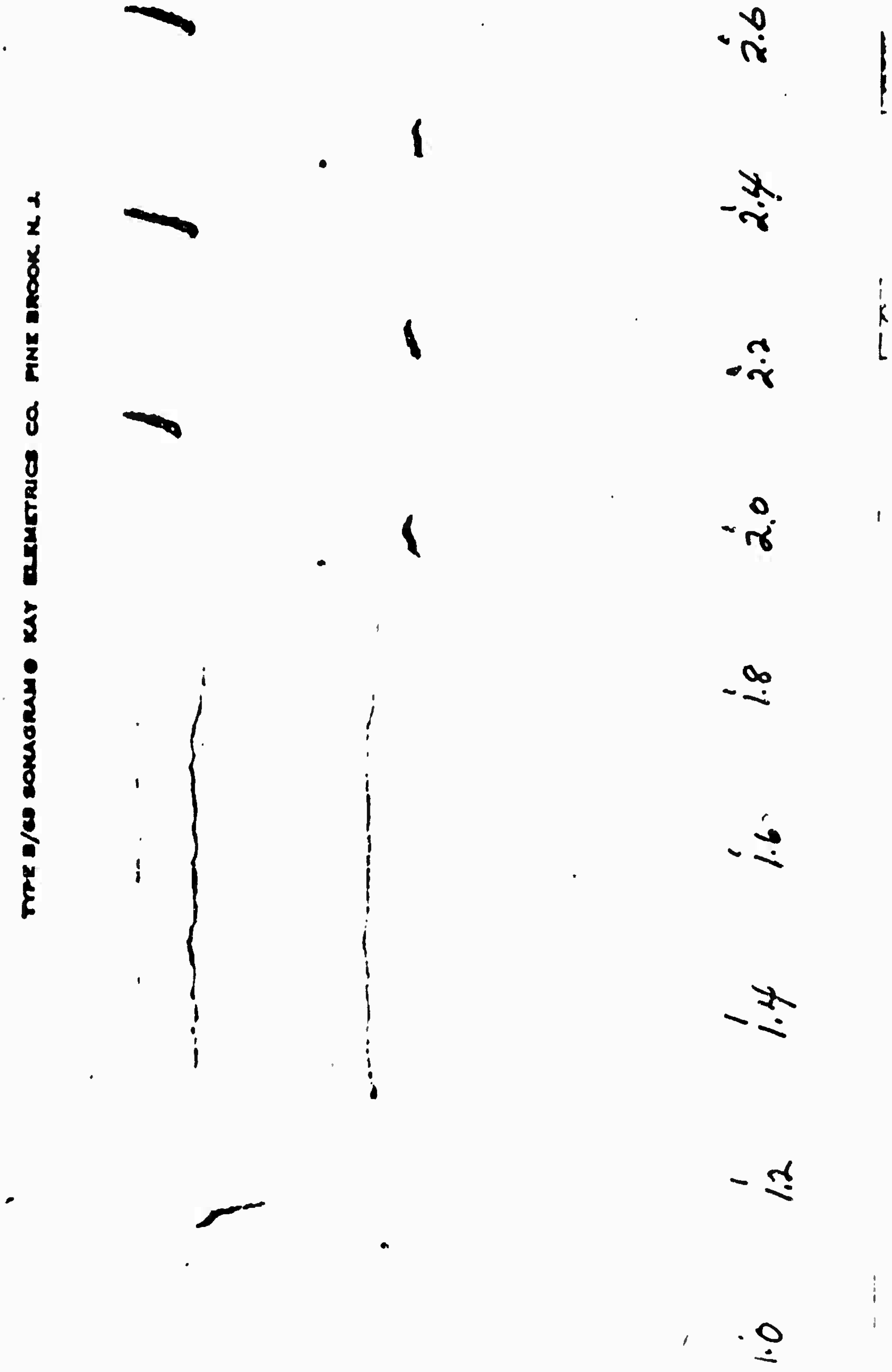
SONG SPARROW (*Melospiza melodia*)



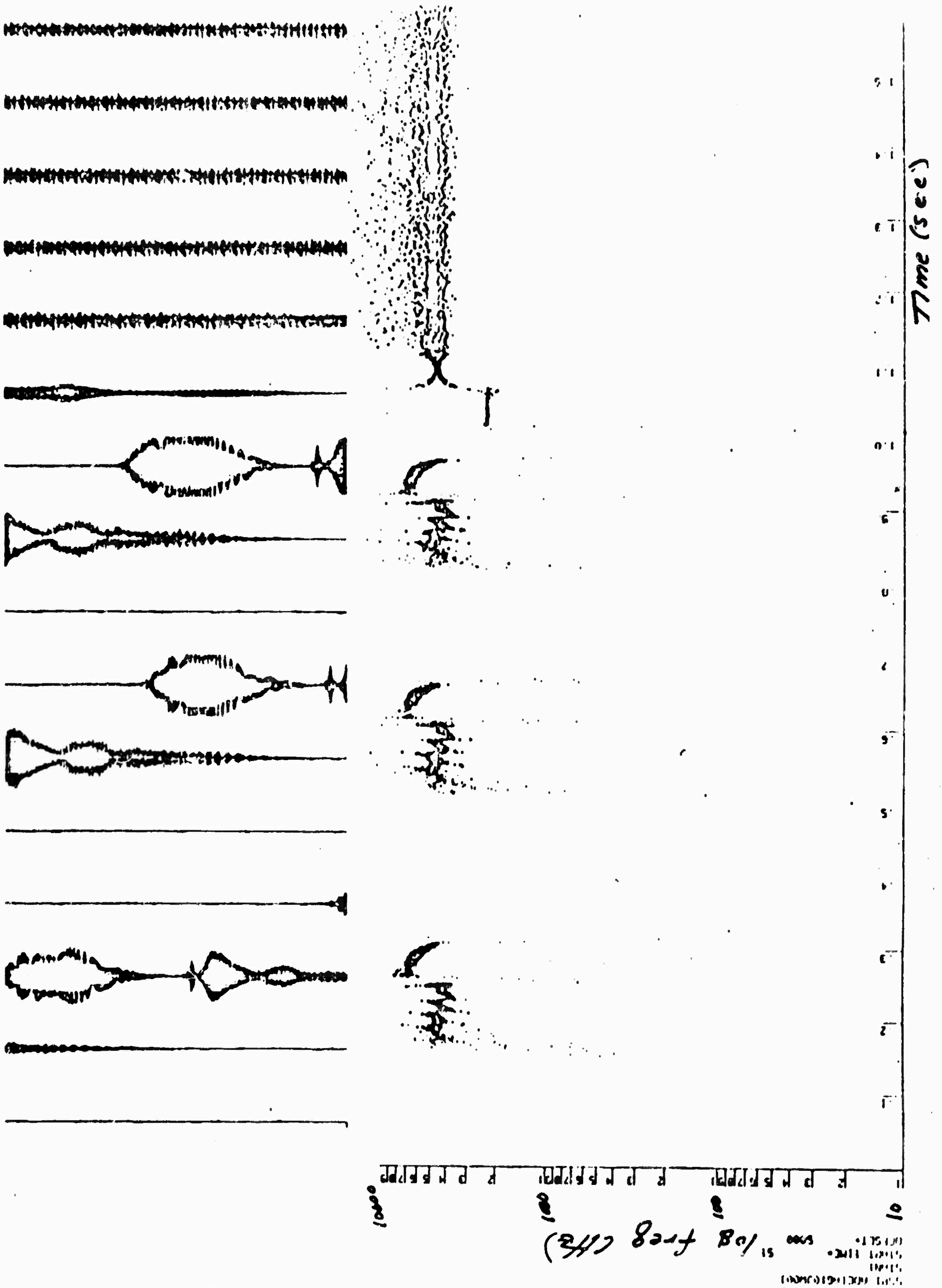
SONG SPARROW (*Melospiza melodia*)



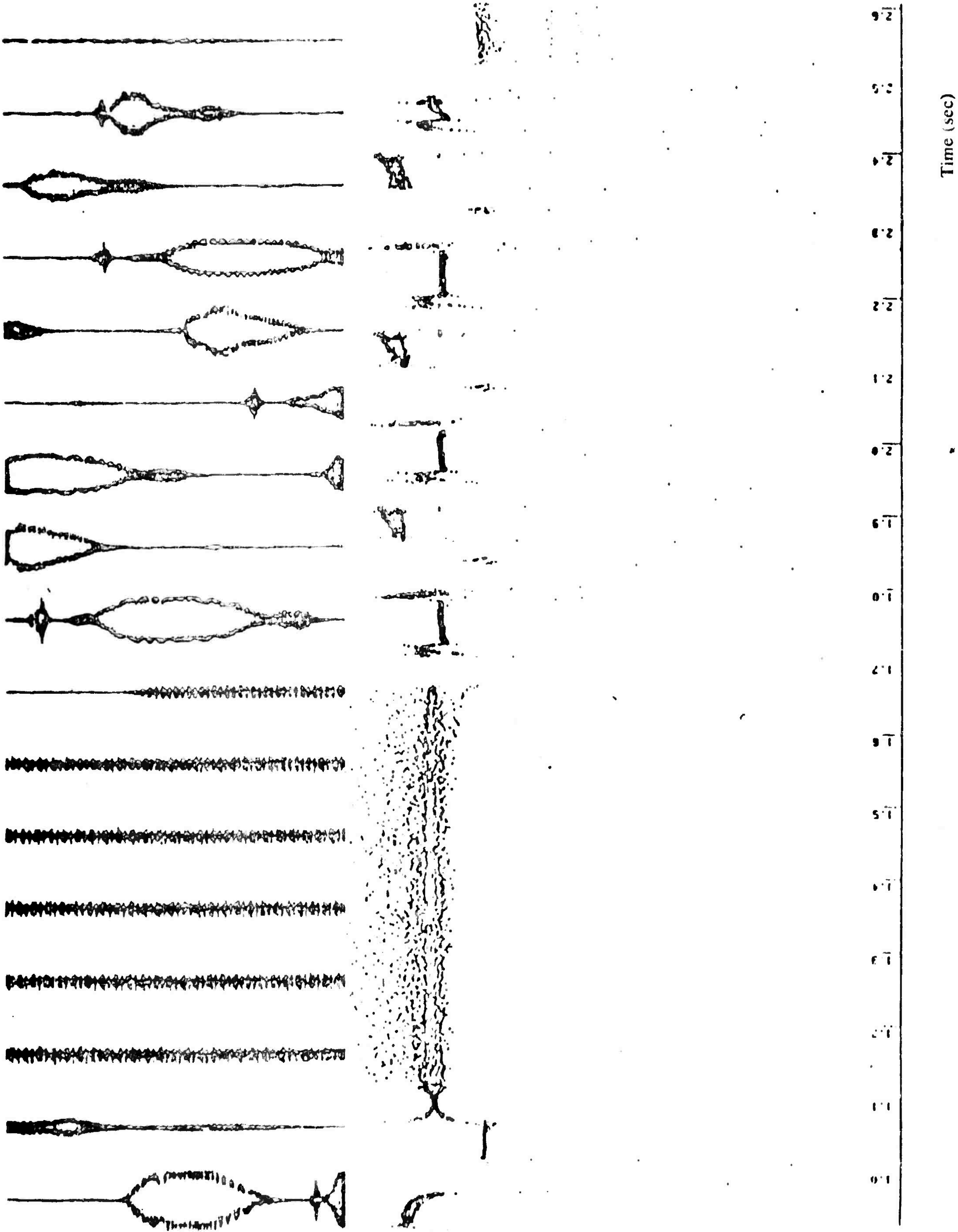
SONG SPARROW (*Melospiza melodia*)



SONG SPARROW (*Melospiza melodia*)



SONG SPARROW (*Melospiza melodia*)



SONG SPARROW (*Melospiza melodia*)

CONCLUDING COMMENTS

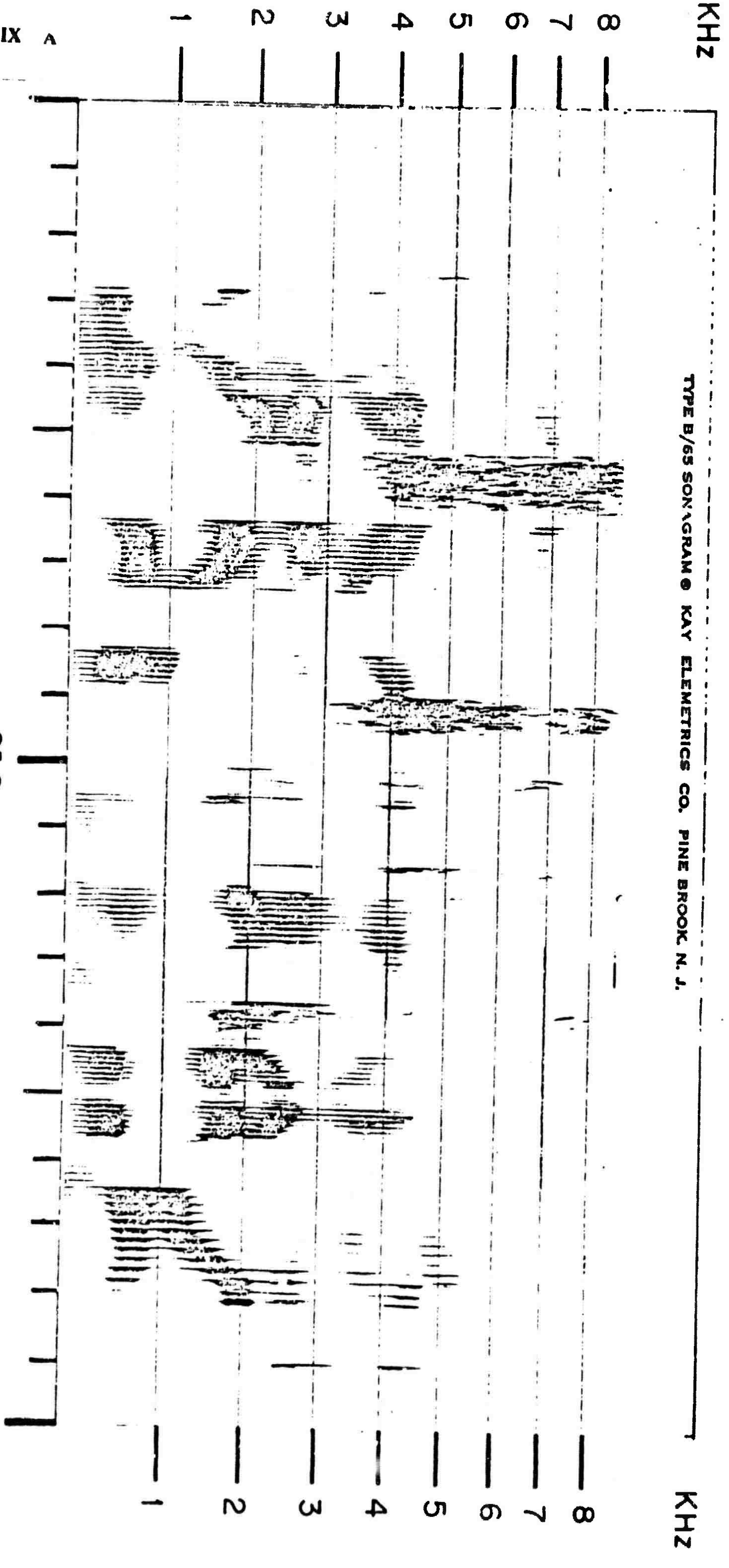
As we have previously stated, the time-domain techniques presented here are, of course, generalizable to any waveform. In light of both the theoretical reasons and empirical results previously discussed, we recommend that these time-domain techniques are most suited to studying waveforms where quickly changing or transient phenomena and/or precise temporal measures are of primary interest.

Despite the historical prevalence and, in fact, nearly exclusive use of frequency-domain analyses for signal waveforms, we have attempted to demonstrate that certain kinds of useful and heretofore untapped information are uniquely available in the time-domain. We have chosen human speech as the chief vehicle for this demonstration, largely because we know much more about the information borne by these complex waveforms than we generally know about most others. As a means for exploring these techniques, we have addressed in detail certain challenging problems in the areas of speech characterization, segmentation, and allophone differentiation. In addition, we have briefly examined other kinds of acoustic waveforms amenable to such analysis methods. It appears to us that a great deal more information in the time-domain remains to be explored, and potentially, quite profitably so.

We strongly recommend however, that frequency-domain analyses and time-domain analyses should be used in a complementary fashion. Despite their redundancy, each of these domains best conveys different and essential characteristics of complex waveforms. We know there are a number of known biological mechanisms of the sensory modalities which are capable of responding to very short term or transient phenomena, as well as others which respond to long term stimuli. In addition, we know that a complex combination of steady states and transients are found in many information-bearing waveforms: human speech, other animal vocalizations, music, biomedical measures such as the EEG and EKG, recordings of physical state changes in inorganic materials, geological seismic recordings, various means of high-frequency communications, and many more.

We feel that the use and understanding of both frequency-domain techniques and time-domain techniques, in conjunction with each other, will lead to better analysis and characterization of complex waveforms, and probably even to improved synthesis of such waveforms as well.

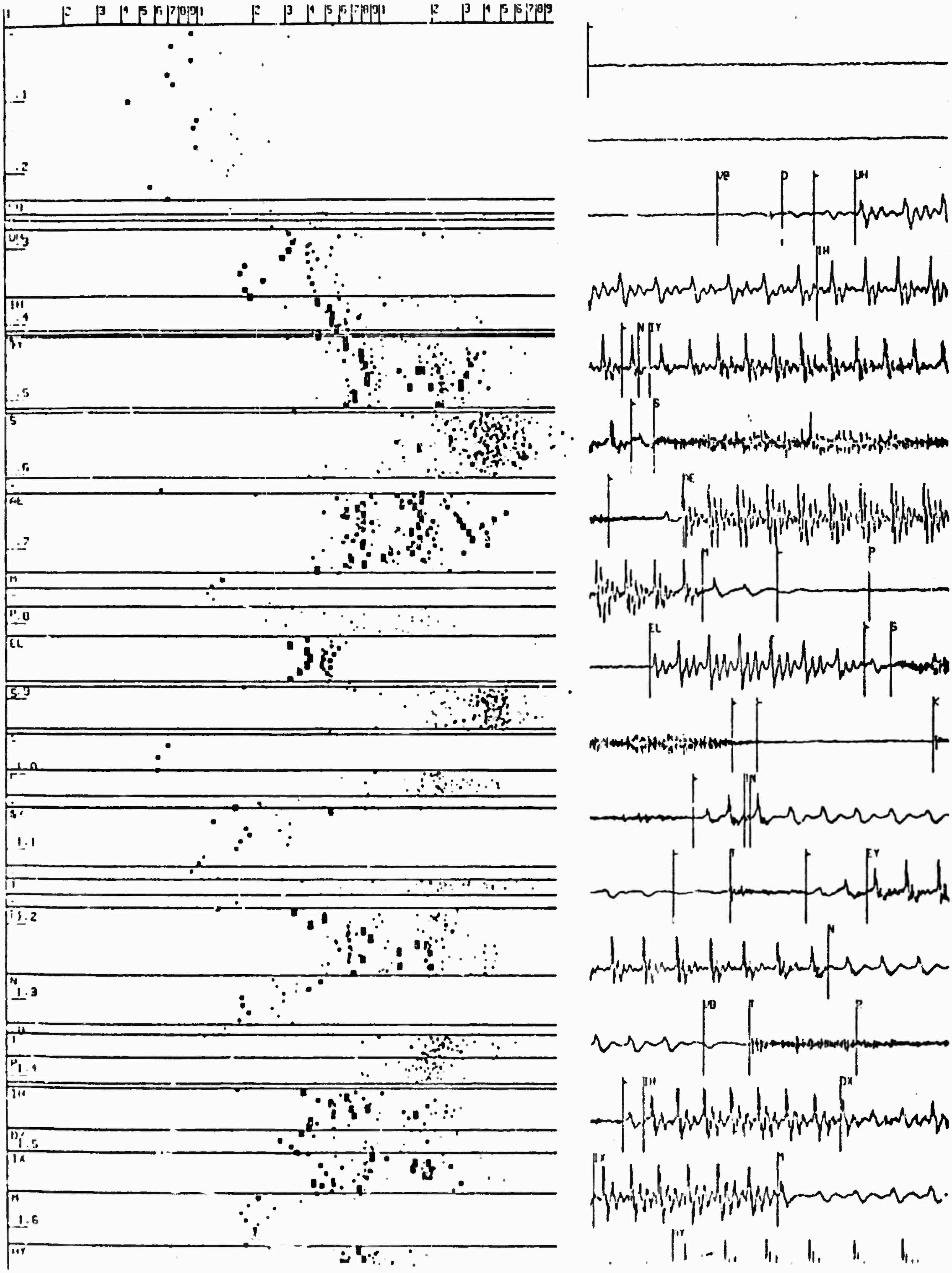
B35: DO ANY SAMPLES CONTAIN TRIDYMITE -- 4a

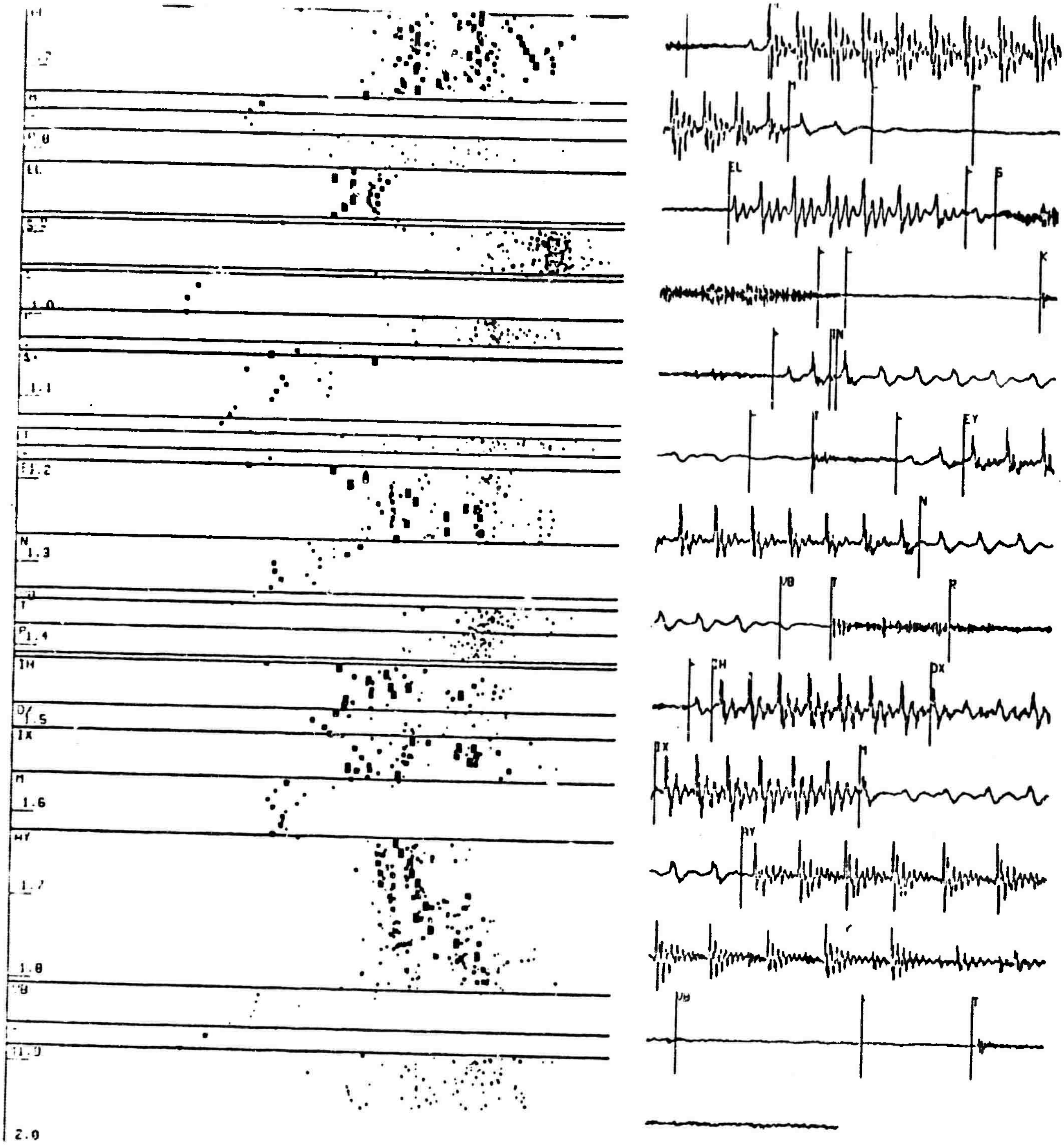


DO ANY SAMPLES CONTAIN TRIDYMITE

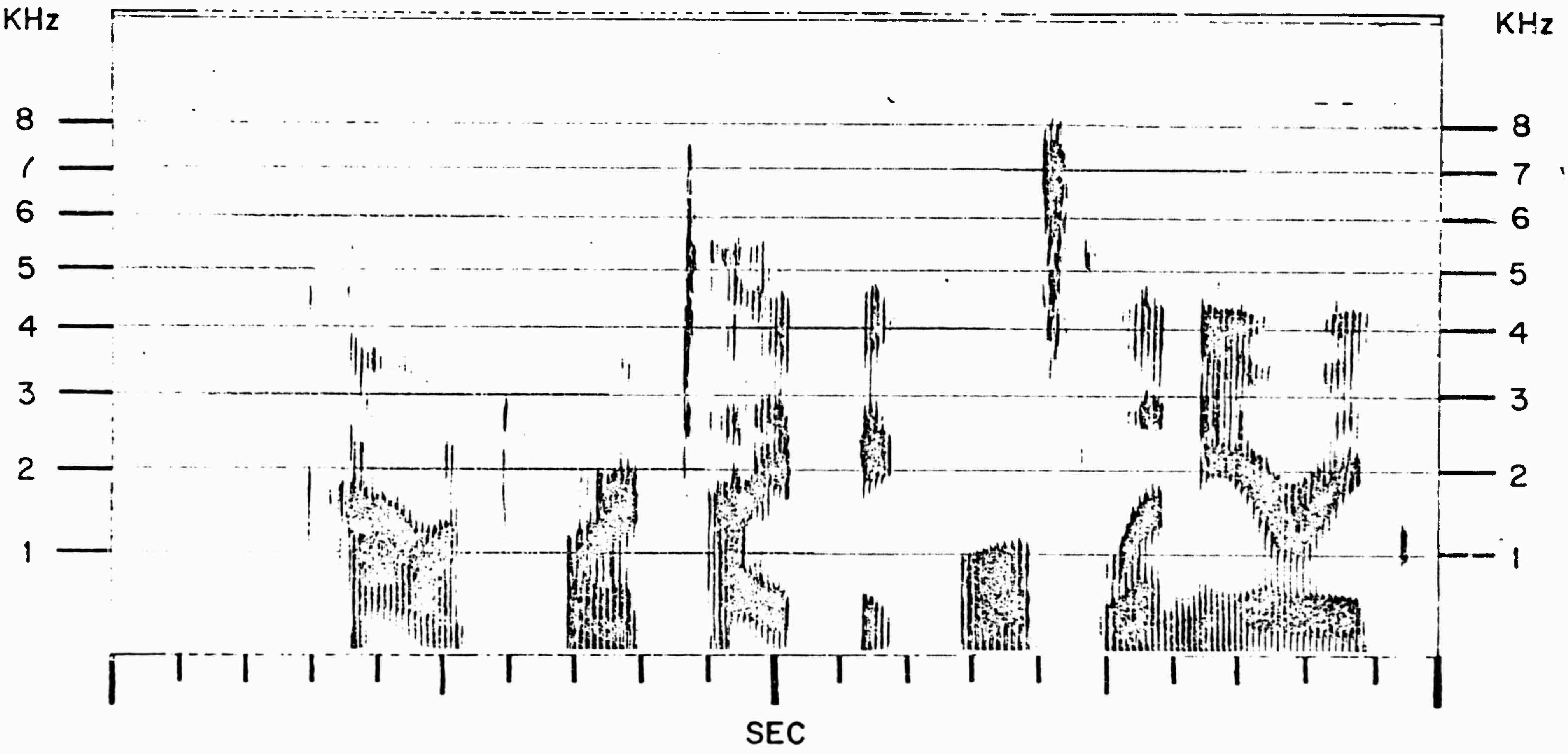
```
D35:**DO ANY SAMPLES CONTAIN TRIDYMITE
LINZO,ADC1AG10JW001
STANZ
START TIME=0
OFFSET=0
```

Log Cycle Frequency (Hz)



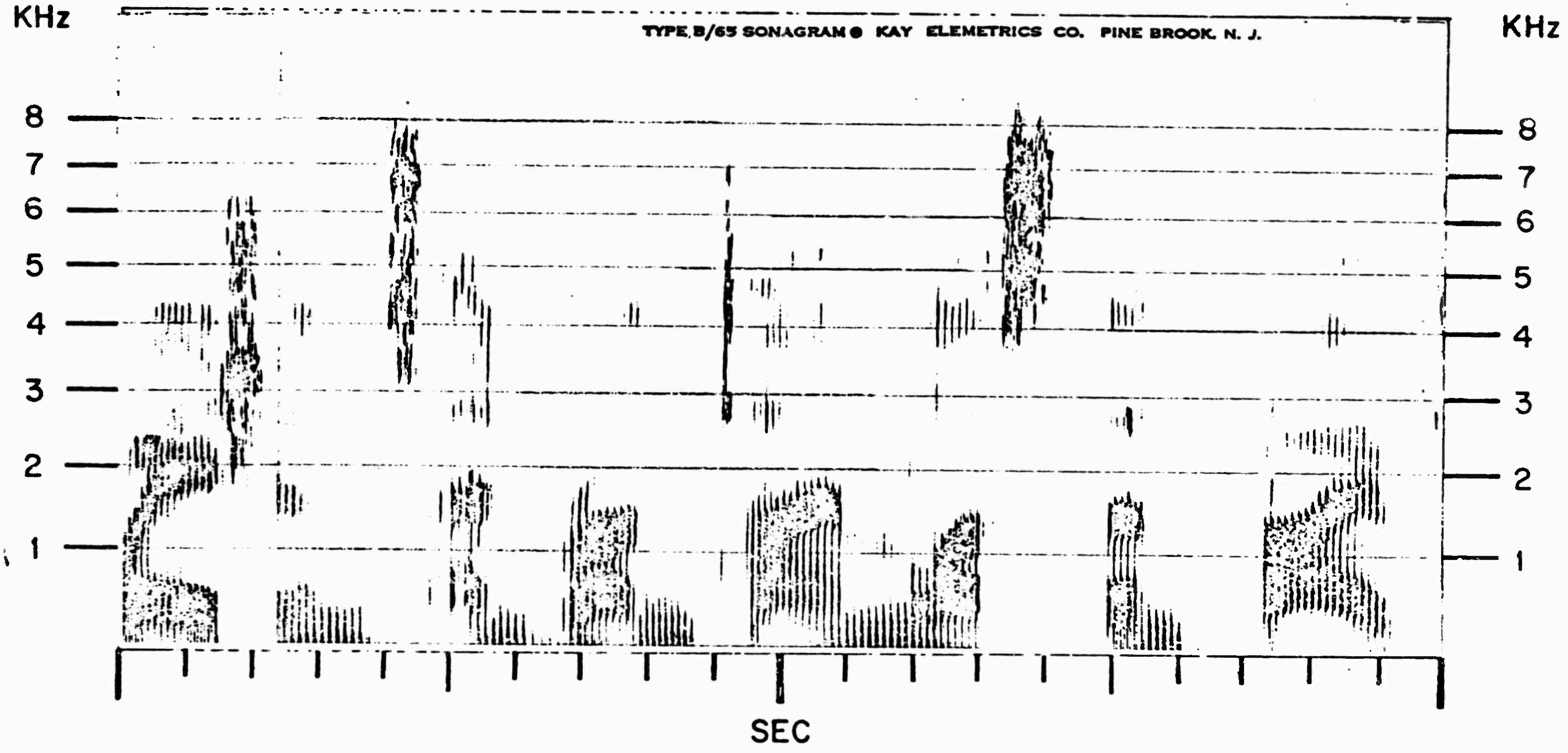

M
0.8
EL
S.9
1.0
1.1
1.2
N
1.3
T
P
1.4
IH
1.5
IX
M
1.6
HY
1.7
1.8
VB
1.9
2.0
EL
S
K
IN
EY
N
VB
T
P
CH
DX
DX
M
HY
VB
T

D7: COUNT WHERE TYPE EQUALS LINEAR EQUATIONS AND RUNTIME LESS THAN F:VE SIX -- 8a



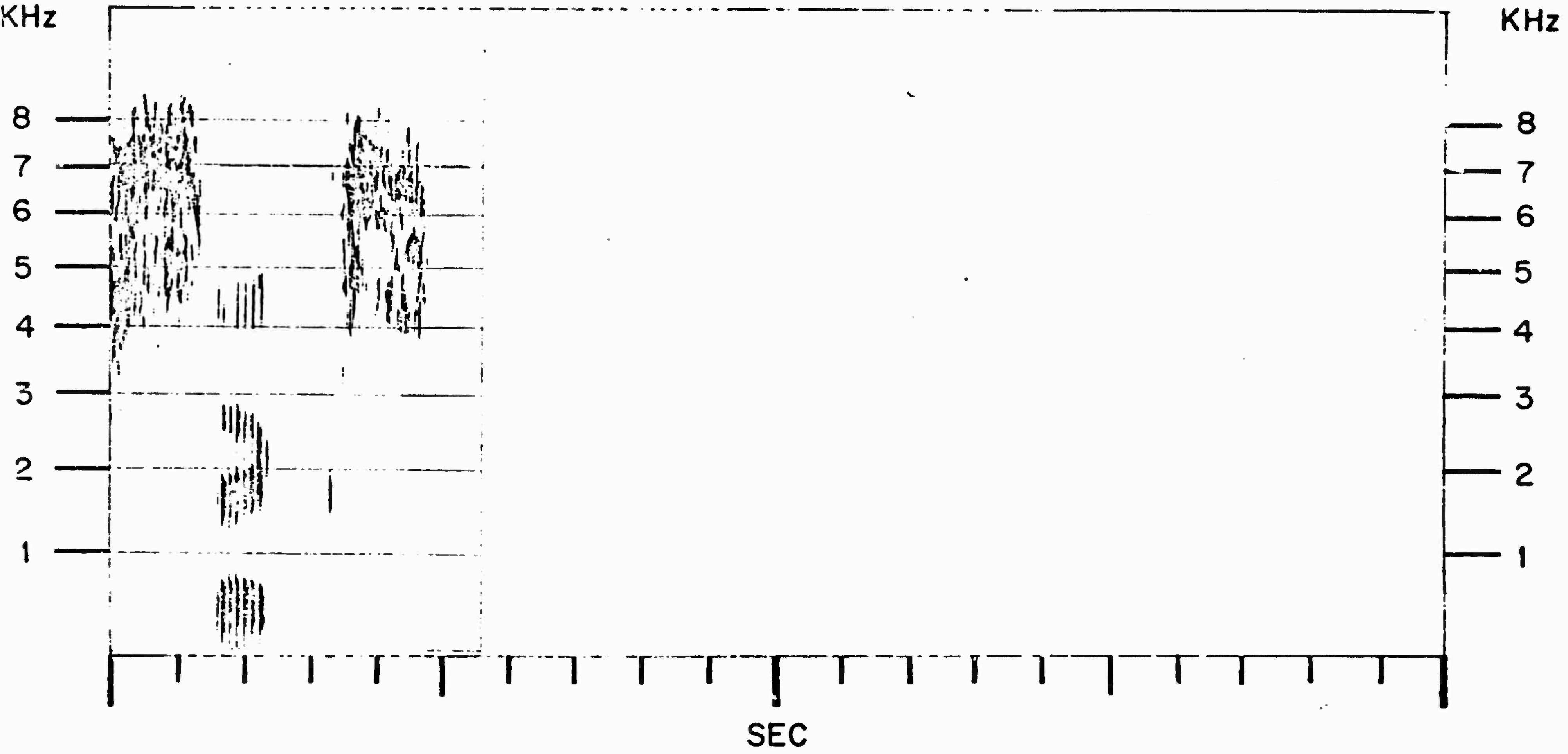
COUNT WHERE TYPE EQUALS LINEAR EQU

D7: COUNT WHERE TYPE EQUALS LINEAR EQUATIONS AND RUNTIME LESS THAN FIVE SIX -- 8b



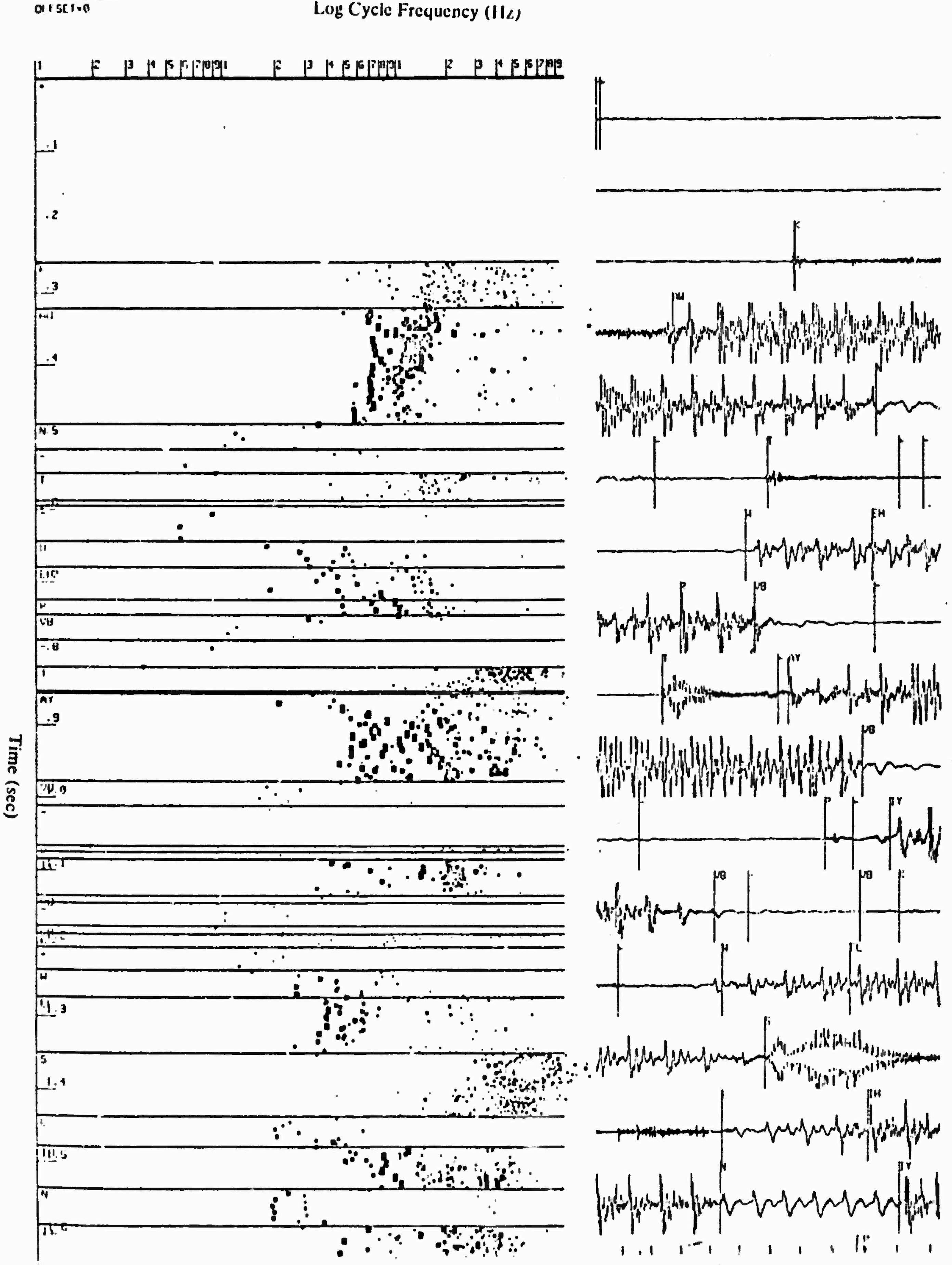
UATIONS AND RUNTIME LESS THAN FIVE S

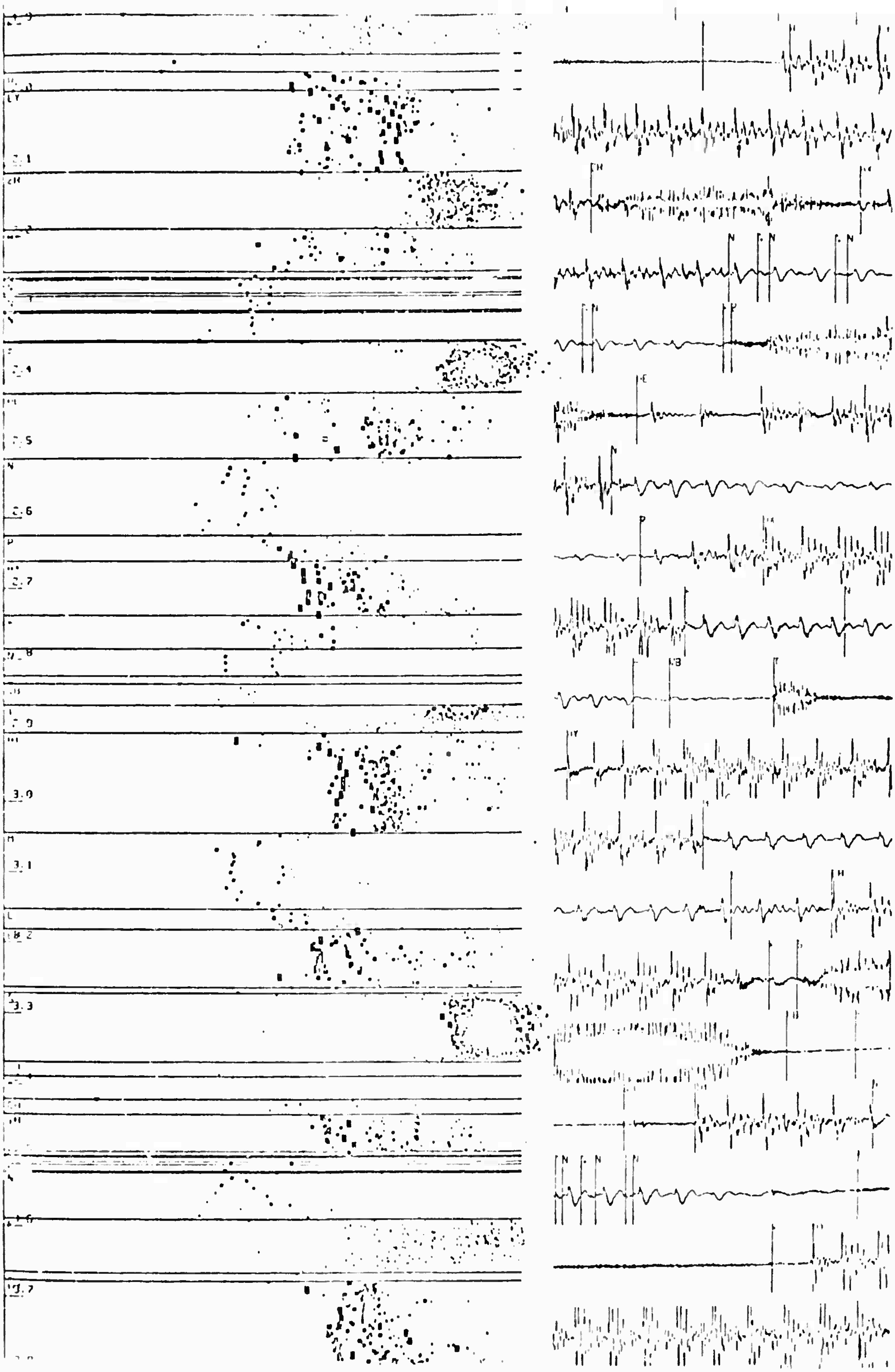
D7: COUNT WHERE TYPE EQUALS LINEAR EQUATIONS AND RUNTIME LESS THAN FIVE SIX -- 8c

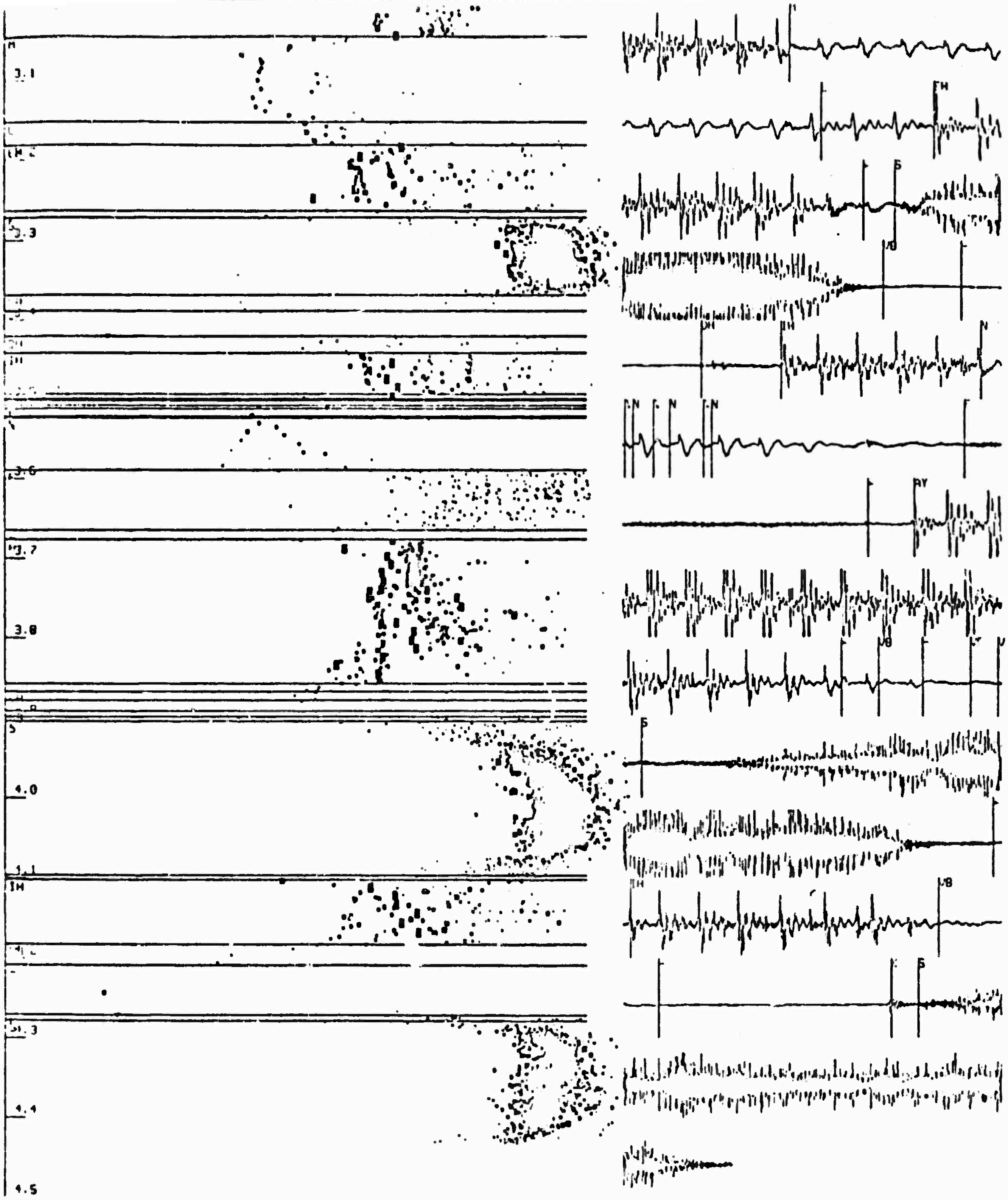


SIX

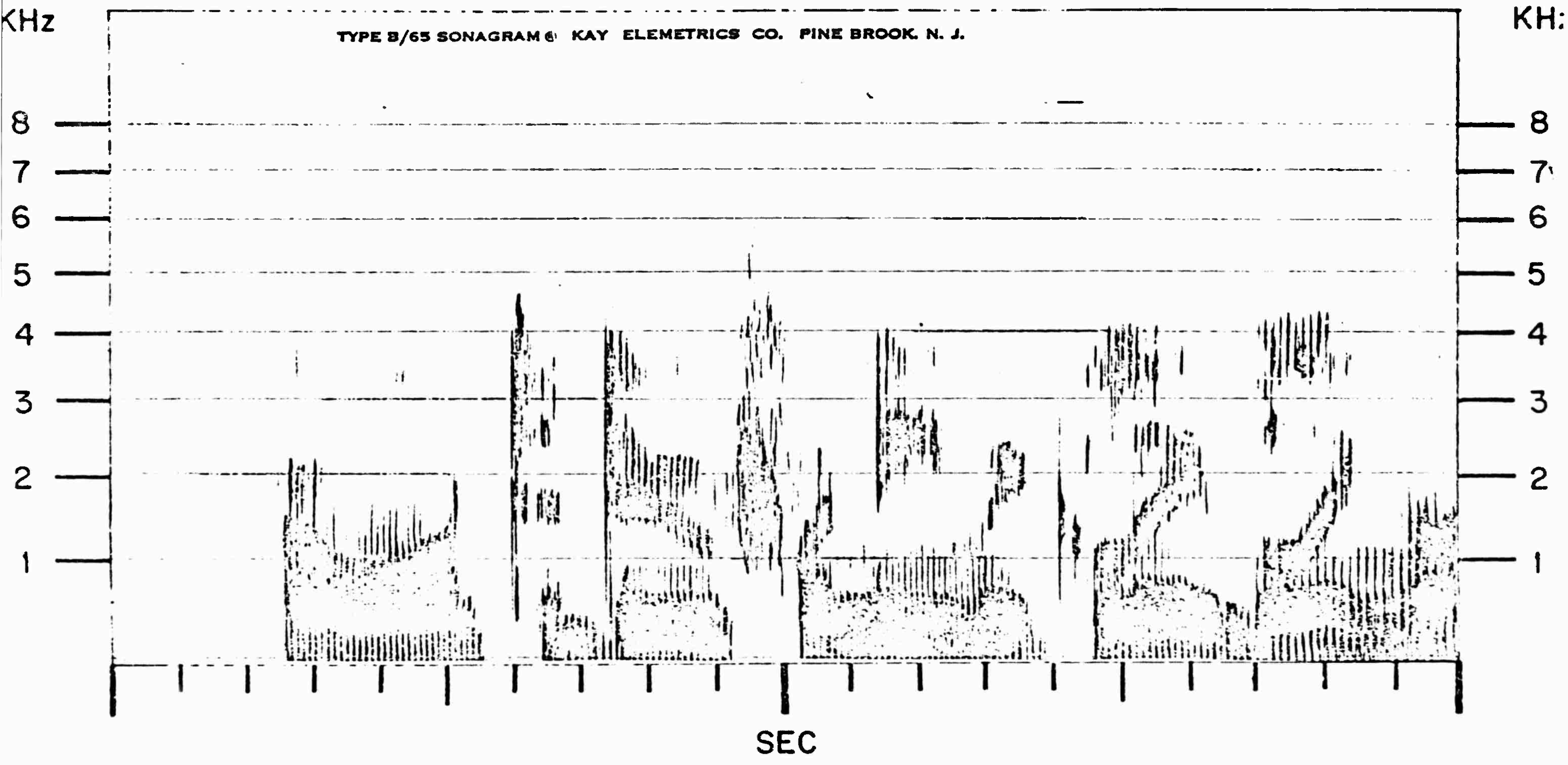
```
D7:=DCOUNT WHERE TYPE EQUALS LINEAR EQUATIONS AND RUNTIME LESS THAN FIVE SIX
LIN20.6OC1AG10JH001
STAN)
START TIME=0
OFFSET=0
```




2.1
2.4
2.5
2.6
2.7
3.0
3.1
3.3
N
P
E
H

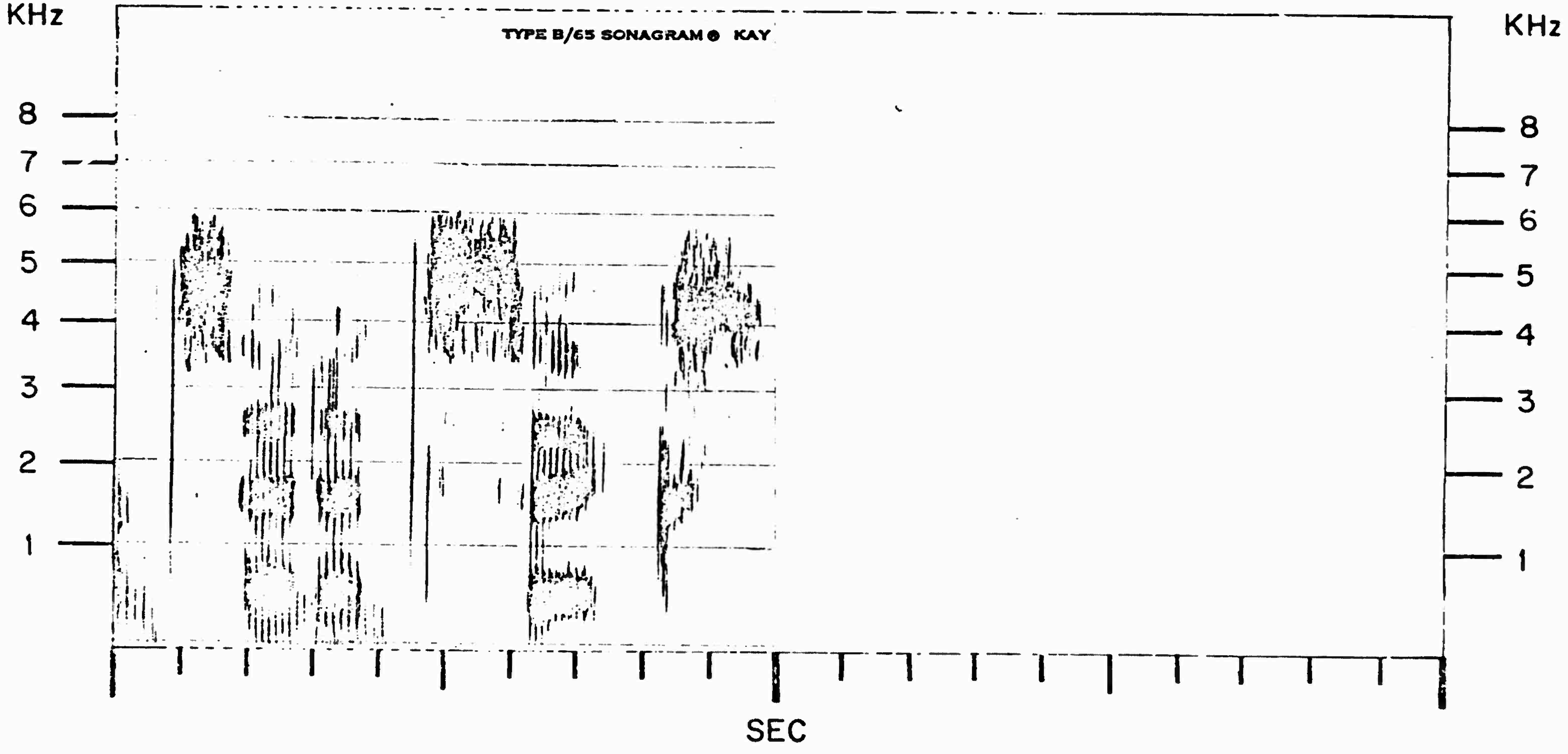

3.1
3.3
3.6
3.7
3.8
4.0
4.4
4.5
H
L
S
B
N
AY

LS1: I WANT TO DO PHONEMIC LABELLING ON SENTENCE SIX -- 12a



I WANT TO DO PHONEMIC LABELING 0

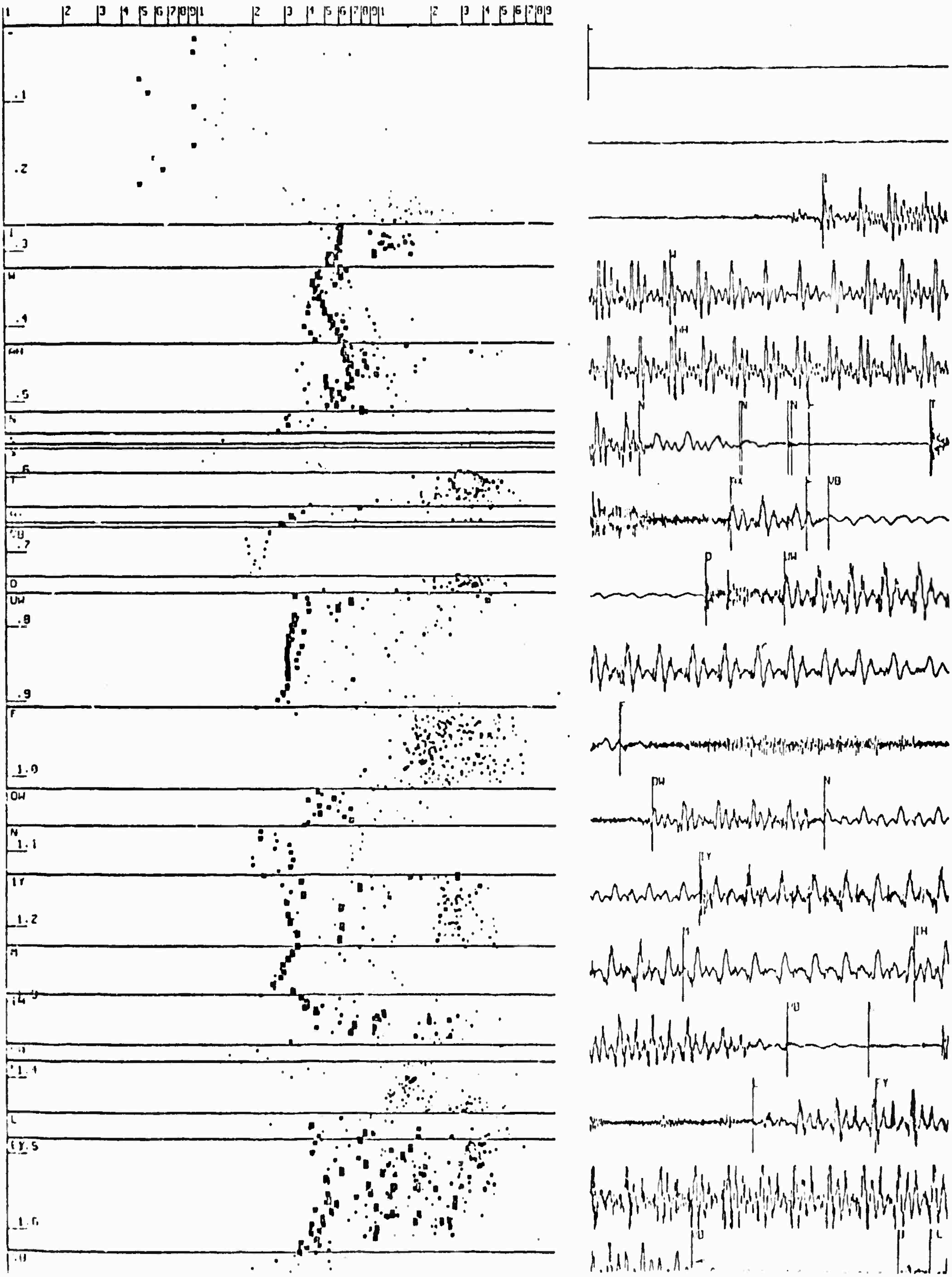
LS1: I WANT TO DO PHONEMIC LABELLING ON SENTENCE SIX -- 12b

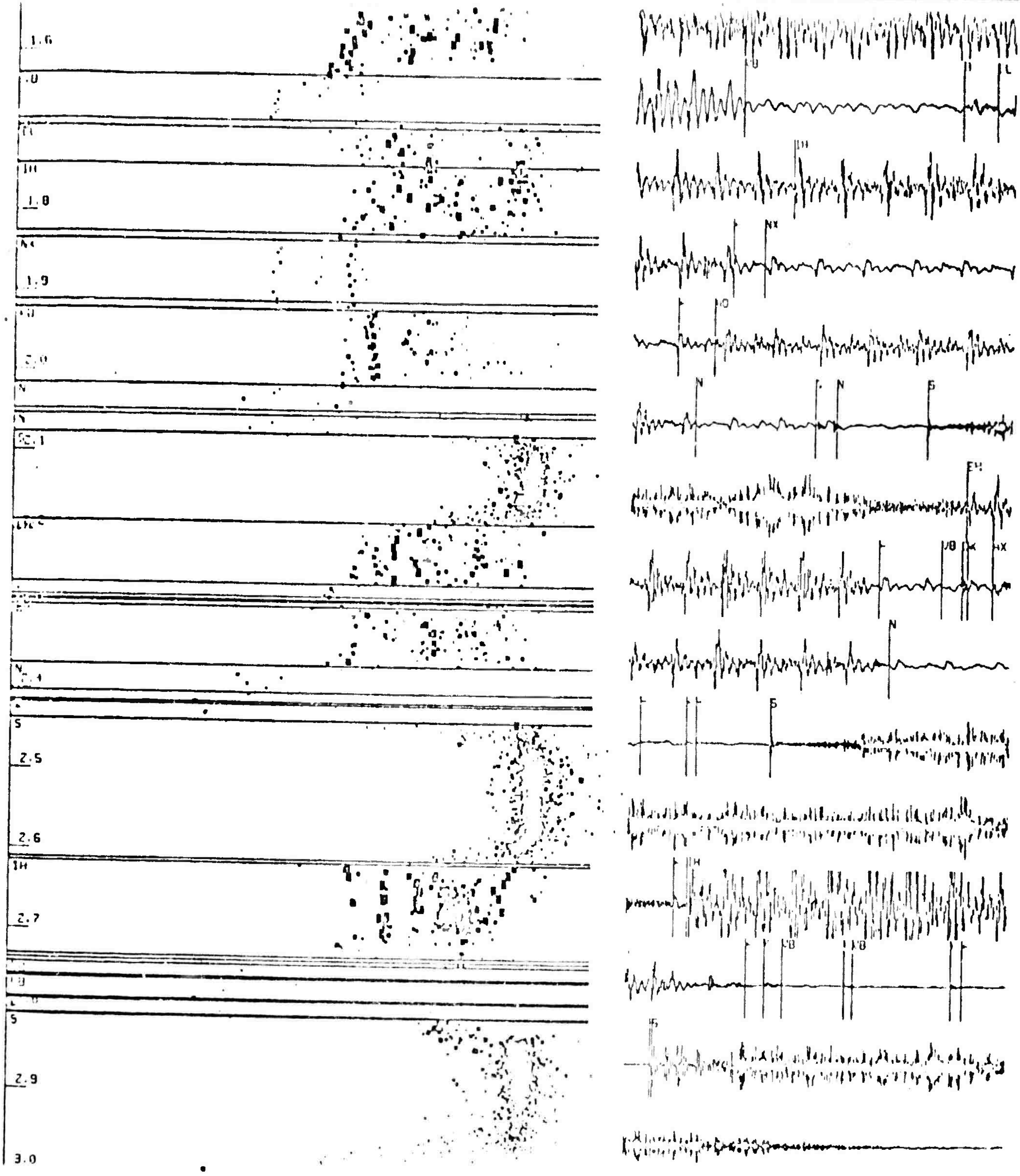


ON SENTENCE SIX

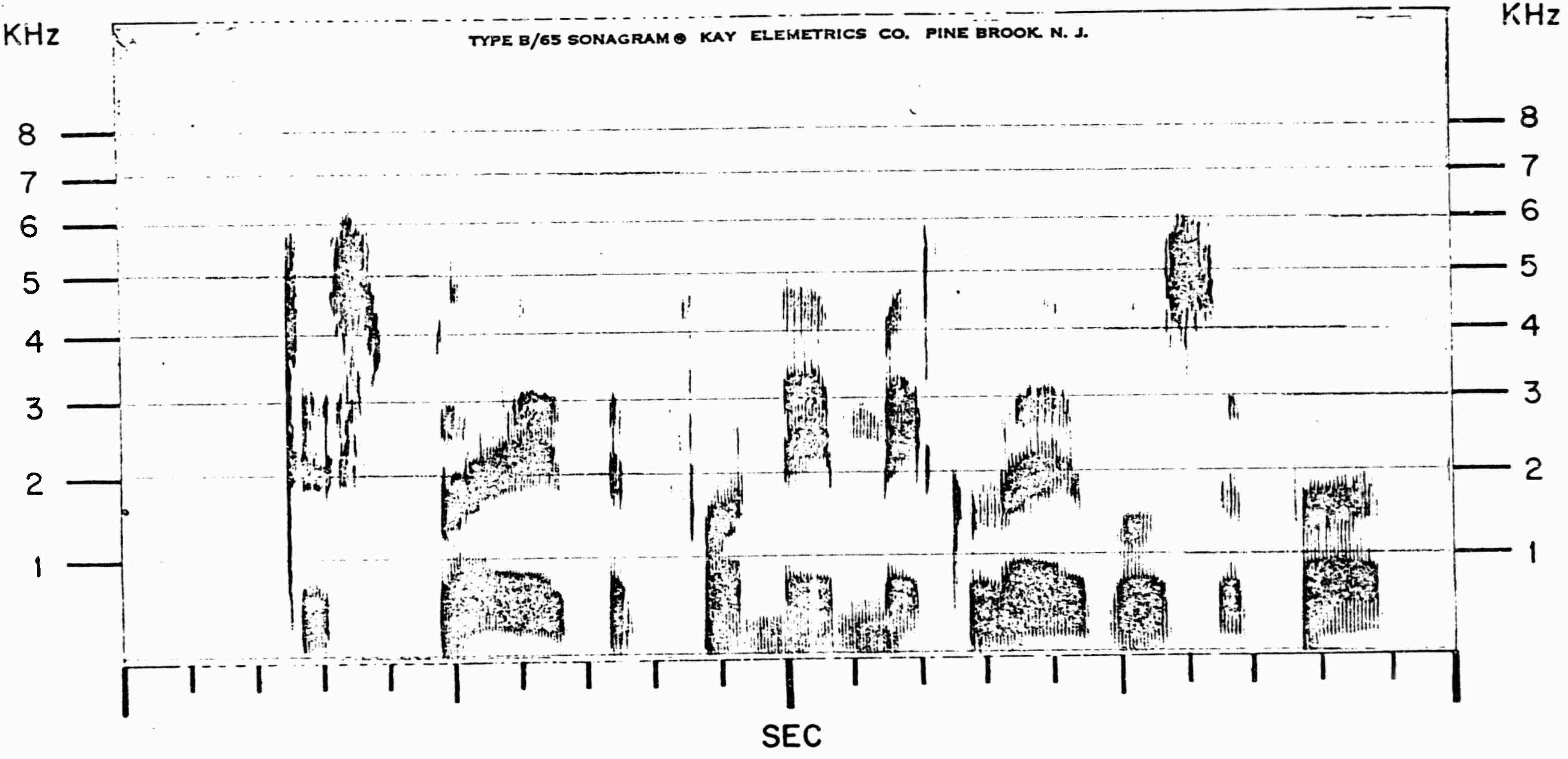
```
LS1:=121 WANT TO DO PHONEMIC LABELING ON SENTENCE SIX
LINZ0.ADC10610JW001
STAN1
START TIME=      2749
OFFSET=          0
```

Log Cycle Frequency (Hz)



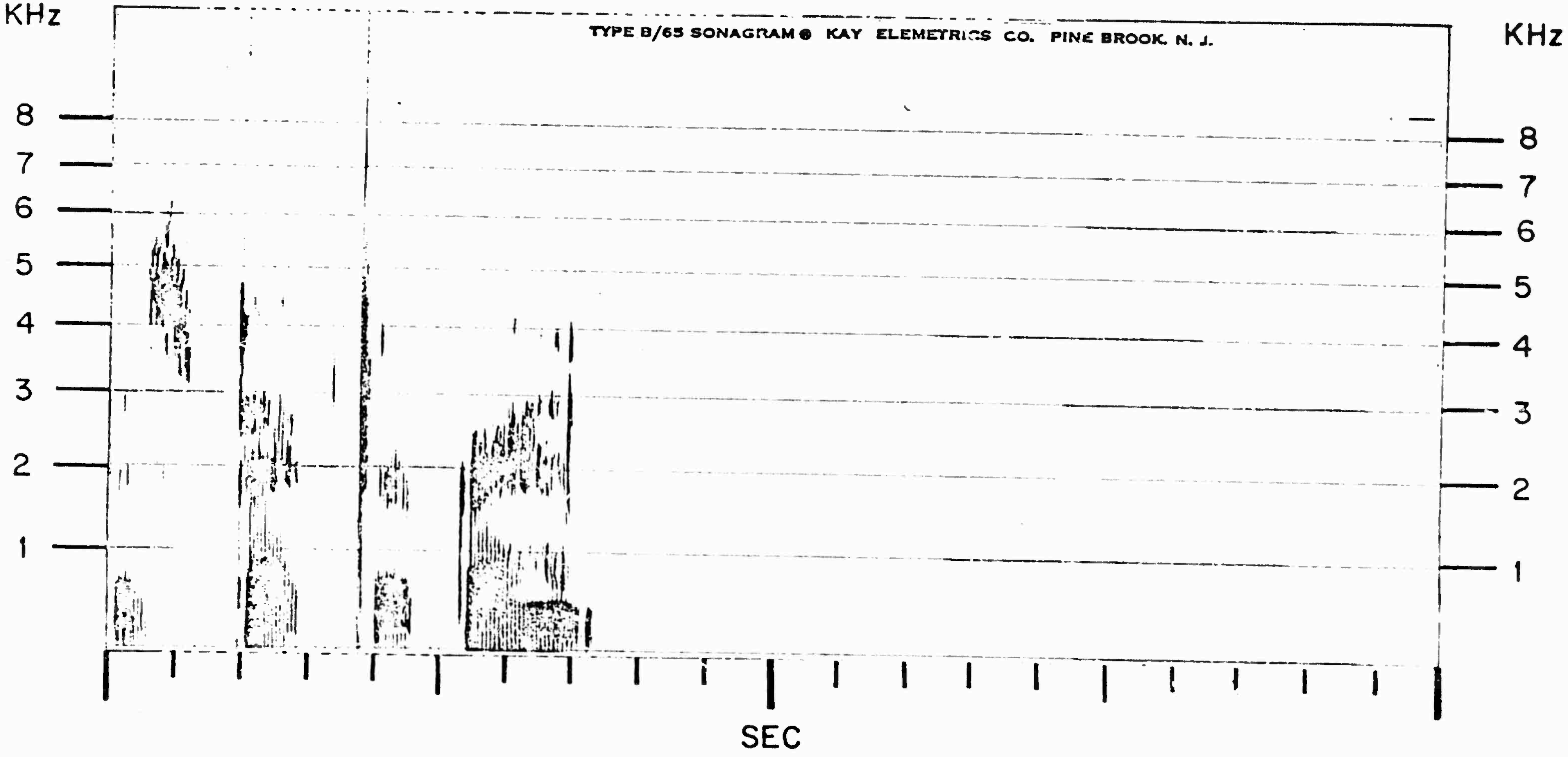

1.6
1.8
1.9
2.0
2.5
2.6
TH
2.7
S
2.9
3.0
NX
N
EH
N
S

LM13: DISPLAY THE PHONEMIC LABELS ABOVE THE SPECTROGRAM -- 15a



DISPLAY THE PHONEMIC LABELS ABOVE TH

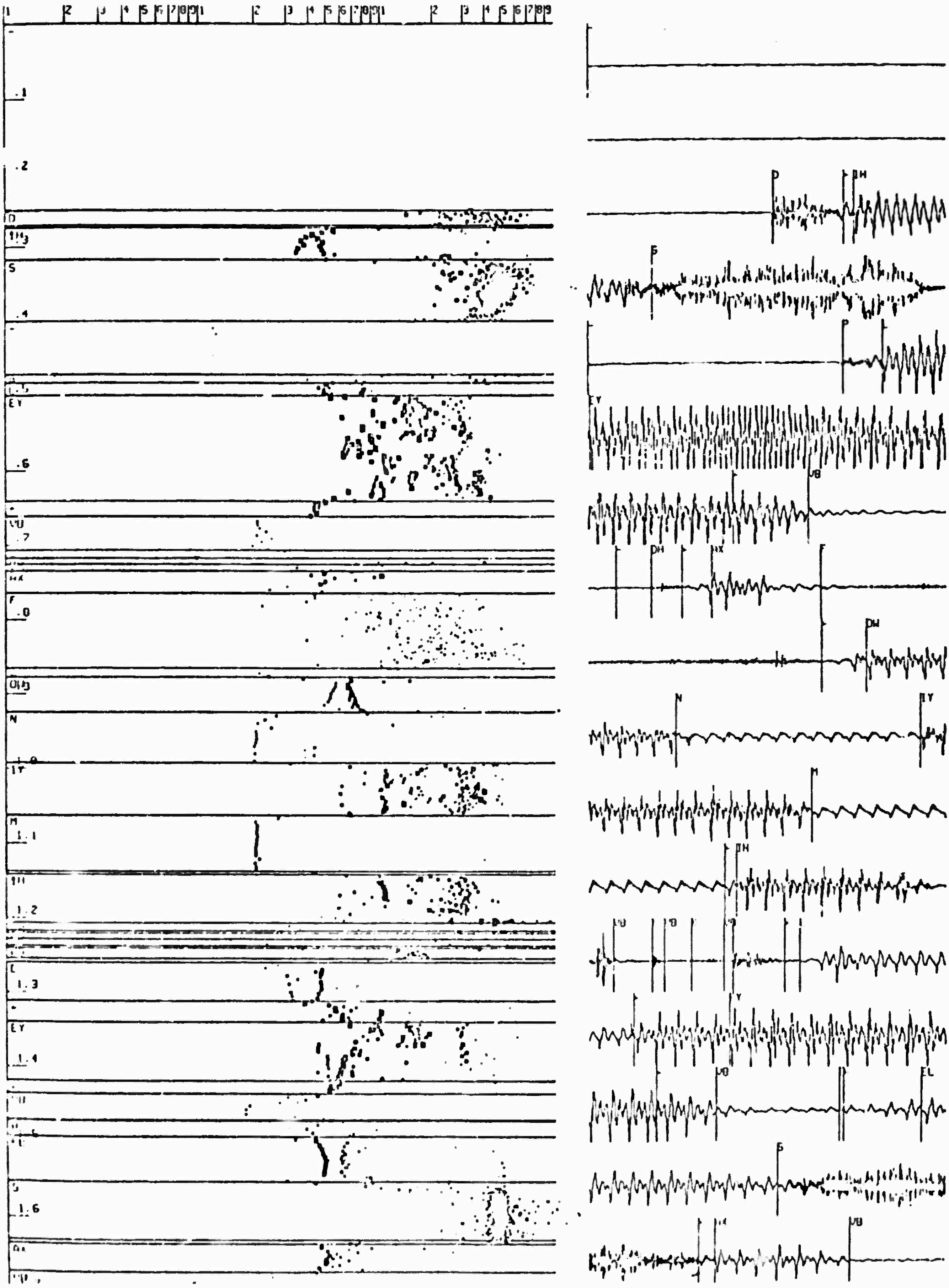
LM13: DISPLAY THE PHONEMIC LABELS ABOVE THE SPECTROGRAM -- 15b



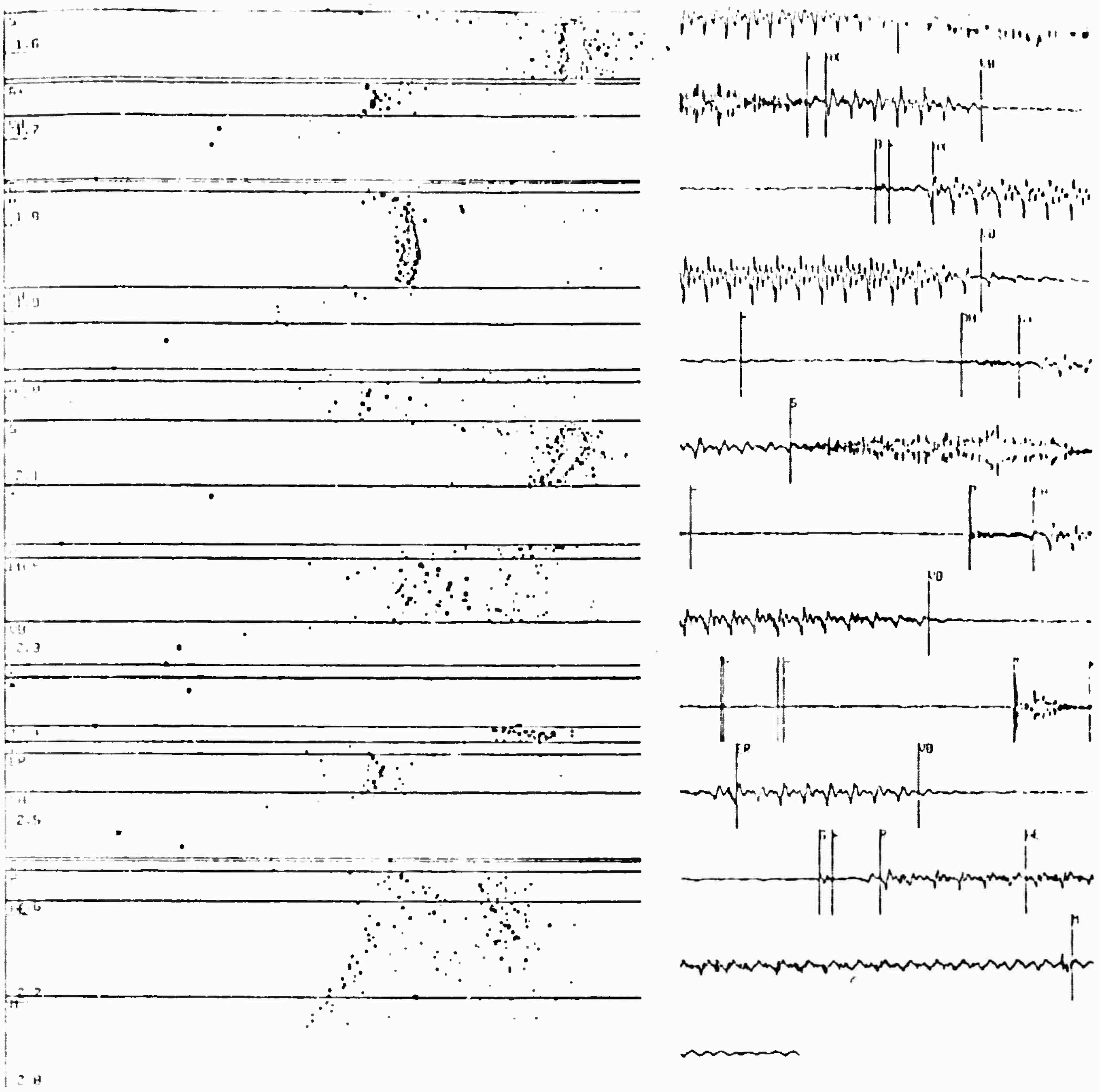
THE SPECTROGRAM

```
LMI3:=ISDISPLAY THE PHONEMIC LABELS ABOVE THE SPECTROGRAM
LINZ0.ADC1AG10JW001
STANI
START TIME=      4474
OFFSET=          0
```

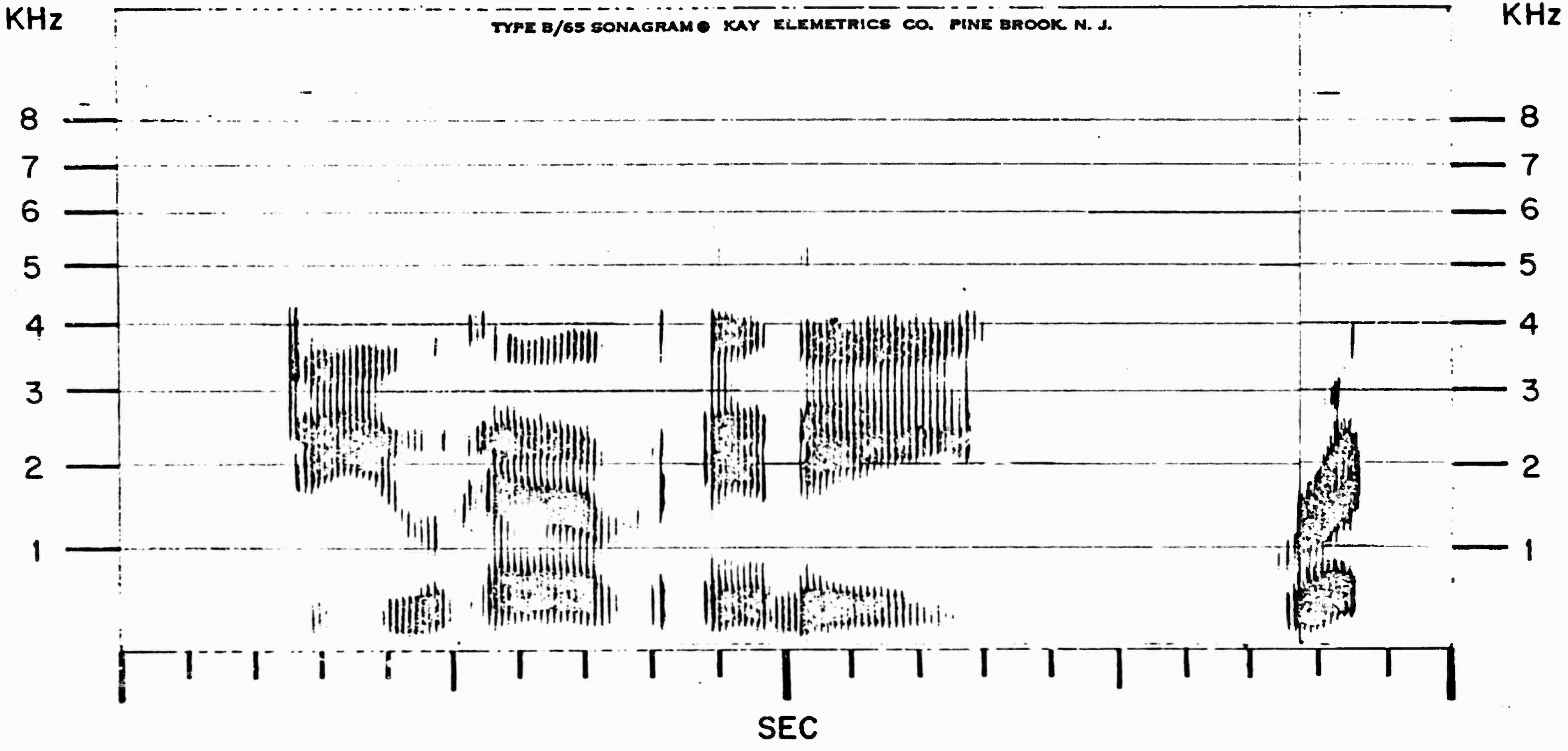
Log Cycle Frequency (Hz)



Page 1

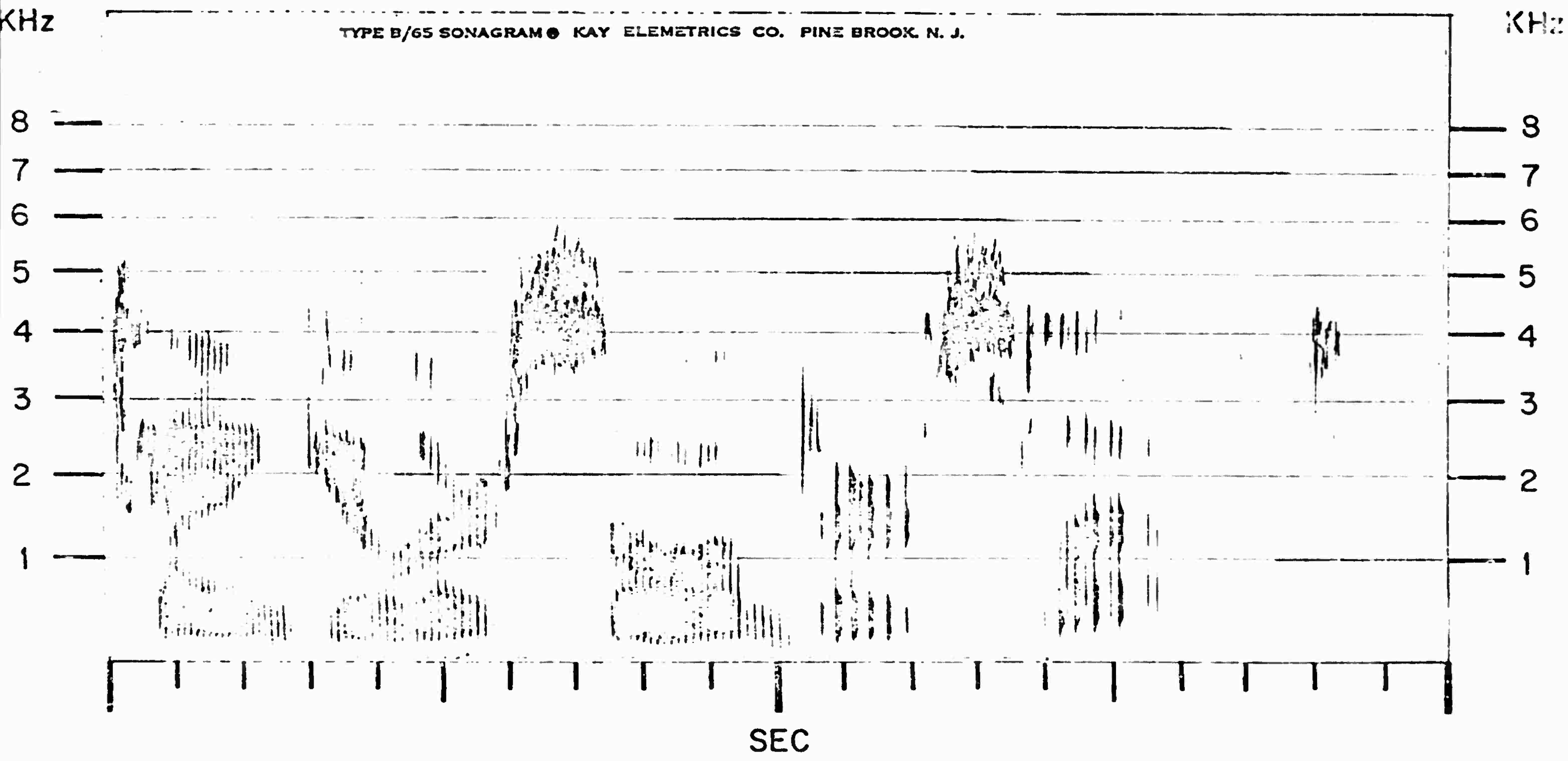


RB7: DO YOU HA√E ANY RECTANGULAR CYLINDERS LEFT -- 21a



DO YOU HAVE ANY REC

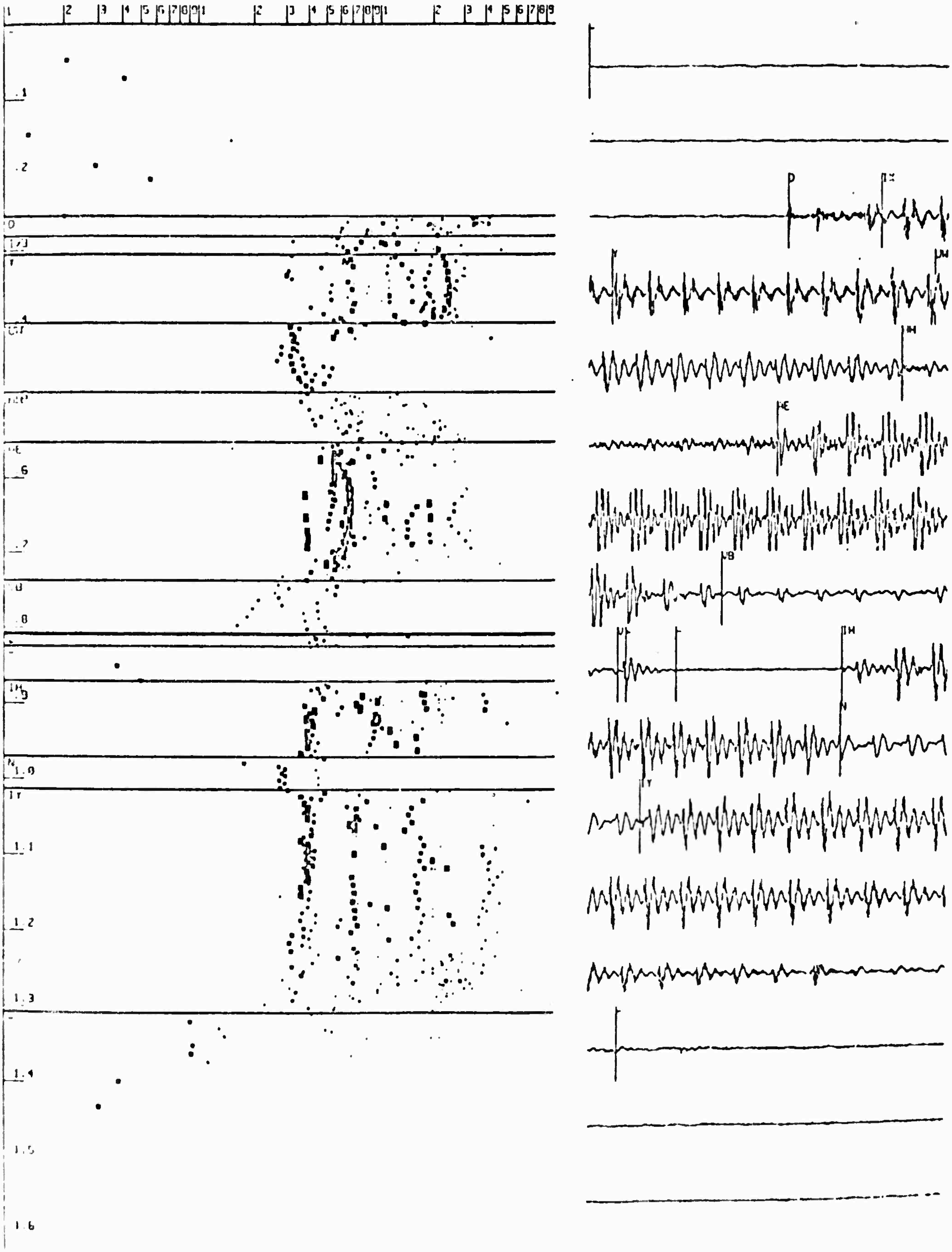
R37: DO YOU HAVE ANY RECTANGULAR CYLINDERS LEFT -- 21b

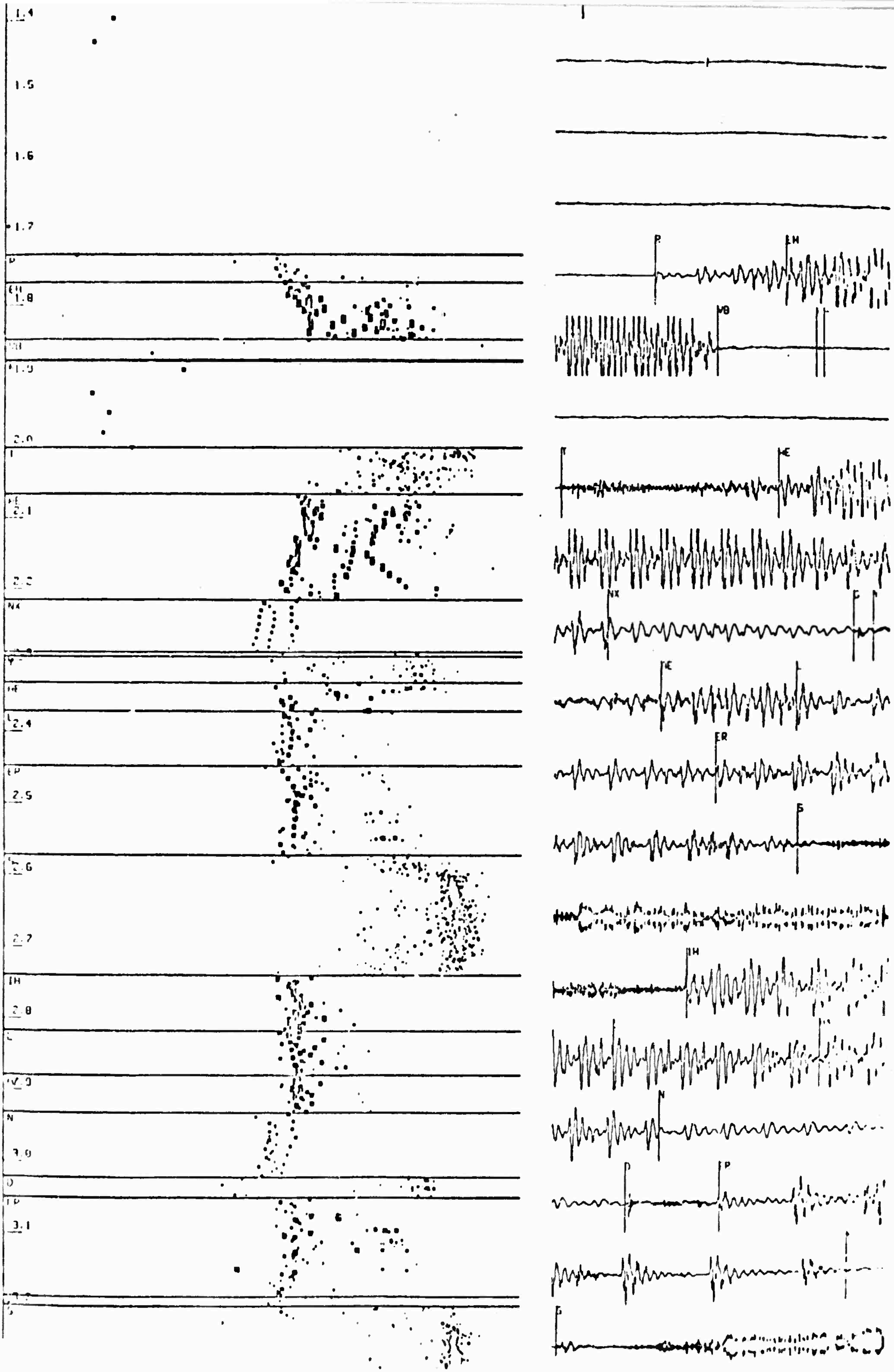


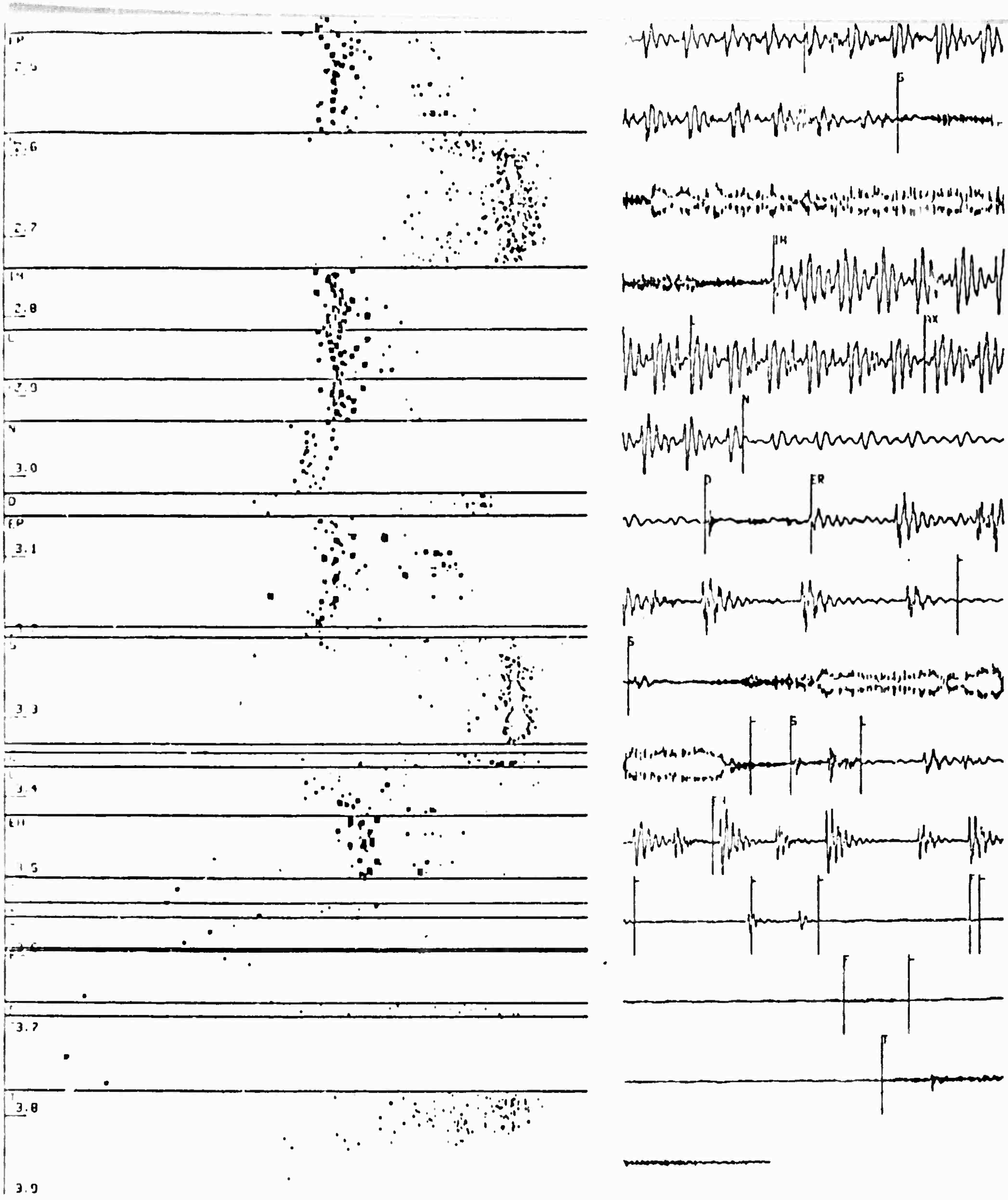
TANGULAR CYLINDERS LEFT

```
PROMPT=DO YOU HAVE ANY RECTANGULAR CYLINDERS LEFT
FILE= [illegible]OC1AG10JK001
STAN2
START TIME=      1209
OFFSET=         0
```

Log Cycle Frequency (Hz)




1.4
1.5
1.6
1.7
P
EH
1.8
VB
1.9
2.0
T
IE
2.1
2.2
NX
Y
IE
L
2.4
EP
2.5
2.6
2.7
IH
2.8
L
2.9
N
3.0
D
EP
3.1
S
P
EH
VB
T
IE
NX
F
IE
L
ER
S
IH
N
D
P
S


2.5
2.6
2.7
2.8
2.9
3.0
D
3.1
3.3
3.4
3.5
3.7
3.8
3.9
N
D
ER
G

APPENDIX B

Sentence Segmentation Results

UTTERANCE #1

1) "Do any samples contain tridymite?"

total # primary boundaries = 33

Absolute Deviation (msec) of Primary Boundaries by Phone Class

PROGRAMS	PAUSES	STOPS	VOWELS	LIQUIDS	FRICATIVES	NASALS	VOICING
IDS	9.2(4)	2.4(7)	3.3(7)	2.1(1)	1.8(2)	1.3(6)	2.3(3)
B	- (0)	10.2(5)	13.3(7)	3.3(2)	7.4(2)	19.0(4)	17.5(2)
C	4.4(2)	6.5(4)	14.9(8)	7.3(1)	0.8(2)	7.3(4)	- (0)
D	5.6(2)	3.7(6)	7.2(8)	0.9(1)	5.7(2)	5.3(5)	3.0(2)
E	8.8(3)	1.5(3)	12.0(8)	9.4(2)	11.5(2)	5.4(4)	5.9(3)

PRIMARY END BOUNDARIES

PROGRAMS	STOPS	FRICATIVES
IDS	3.6(5)	0.1(2)
B	8.5(6)	17.2(1)
C	12.3(5)	8.5(2)
D	5.1(4)	4.3(2)
E	12.3(6)	11.5(2)

PROGRAMS	MISSED BOUNDARIES	EXTRA BOUNDARIES	MISSED PLUS EXTRAS	# SECONDARY BOUNDARIES
IDS	3	4	7	6
B	11	5	16	2
C	12	3	15	1
D	7	6	13	2
E	8	4	12	1

UTTERANCE #2

2) "Count where type equals linear equations and runtime less than five six."

total # primary boundaries = 64

Absolute Deviation (msec) of Primary Boundaries by Phone Class

PROGRAMS	PAUSES	STOPS	VOWELS	LIQUIDS	FRICATIVES	NASALS	VOICING
TDS	0.3(10)	1.2(8)	11.2(13)	4.3(8)	4.4(7)	1.4(7)	0.9(7)
B	14.6(6)	10.3(8)	12.1(16)	9.2(8)	14.5(7)	12.3(7)	18.0(6)
C	8.6(3)	6.6(2)	7.8(15)	7.2(5)	10.6(8)	11.4(6)	6.9(8).
D	6.0(8)	3.0(8)	6.3(15)	5.7(7)	13.6(7)	6.0(6)	12.0(8)
E	10.5(5)	16.1(5)	10.1(15)	11.1(8)	13.8(5)	10.9(6)	4.1(5)

PRIMARY END BOUNDARIES

PROGRAMS	STOPS	FRICATIVES
TDS	4.4(6)	8.6(6)
B	12.5(6)	17.8(7)
C	5.4(7)	5.2(7)
D	4.7(7)	5.9(7)
E	6.4(5)	10.6(7)

PROGRAMS	MISSED BOUNDARIES	EXTRA BOUNDARIES	MISSED PLUS EXTRAS	# SECONDARY BOUNDARIES
TDS	4	8	12	12
B	6	12	18	1
C	19	0	19	2
D	5	18	23	4
E	15	7	22	3

UTTERANCE #3

3) "I want to do phonemic labeling on sentence six."

total # primary boundaries = 38

Absolute Deviation (msec) of Primary Boundaries by Phone Class

PROGRAMS	PAUSES	STOPS	VOWELS	LIQUIDS	FRICATIVES	NASALS	VOICING
TDS	1.3(3)	0.1(5)	3.9(10)	22.8(1)	7.0(4)	1.1(6)	4.5(4)
B	- (0)	10.0(4)	10.6(11)	8.1(3)	5.3(4)	6.3(5)	2.3(3)
C	3.2(2)	13.1(3)	10.8(11)	11.6(2)	11.8(4)	6.4(6)	2.8(4)
D*	5.7(1)	6.2(1)	14.4(6)	33.0(1)	- (0)	12.2(3)	7.1(1)
E	8.0(3)	6.2(3)	15.3(8)	7.4(2)	25.2(4)	13.7(4)	0.8(3)

PRIMARY END BOUNDARIES

PROGRAMS	STOPS	FRICATIVES
TDS	9.8(4)	2.8(3)
B	5.2(5)	2.3(3)
C	7.8(6)	12.8(1)
D*	20.6(3)	12.3(1)
E	5.8(3)	18.9(3)

PROGRAMS	MISSED BOUNDARIES	EXTRA BOUNDARIES	MISSED PLUS EXTRAS	# SECONDARY BOUNDARIES
TDS	5	6	11	7
B	8	10	18	3
C	6	0	6	1
D*	8	8	16	0
E	11	10	21	1

* - only about half of utterance segmented.

UTTERANCE #4

4) "Display the phonemic labels above the spectrogram."

total # primary boundaries = 47

Absolute Deviation (msec) of Primary Boundaries by Phone Class

PROGRAMS	PAUSES	STOPS	VOWELS	LIQUIDS	FRICATIVES	NASALS	VOICING
TDS	0.8(5)	0.1(9)	3.1(10)	9.0(3)	4.9(6)	0.4(3)	1.1(6)
B	6.4(2)	4.1(5)	8.8(11)	5.7(4)	12.4(5)	7.7(3)	8.0(7)
C	20.0(1)	1.6(2)	13.0(6)	- (0)	27.2(3)	9.9(1)	4.5(1)
D*							
E	22.8(3)	14.8(7)	10.4(10)	13.6(3)	18.3(5)	15.9(2)	10.1(5)

PRIMARY BOUNDARIES

PROGRAMS	STOPS	FRICATIVES
TDS	4.9(5)	2.8(7)
B	9.9(6)	8.3(6)
C	- (0)	18.7(3)
D*		
E	17.1(7)	16.0(7)

PROGRAMS	MISSED BOUNDARIES	EXTRA BOUNDARIES	MISSED PLUS EXTRAS	# SECONDARY BOUNDARIES
TDS	5	7	12	11
B	10	6	16	0
C	33	1	34	0
D*				
E	12	2	14	0

* - utterance segmentation results not available

UTTERANCE #5

5) "Do you have any rectangular cylinders left?"

total # primary boundaries = 34

Absolute Deviation (msec) of Primary Boundaries by Phone Class

PROGRAMS	PAUSES	STOPS	VOWELS	LIQUIDS	FRICATIVES	NASALS	VOICING
IDS	5.0(2)	3.0(6)	5.8(9)	7.9(5)	5.4(4)	4.5(3)	2.9(2)
B	10.1(5)	7.7(4)	12.6(10)	5.7(4)	8.9(3)	13.3(3)	- (0)
C	14.0(1)	18.1(2)	17.0(5)	- (0)	16.1(4)	2.5(1)	13.1(1)
D	10.8(3)	7.3(5)	11.4(11)	13.9(5)	11.4(4)	10.9(3)	11.3(1)
E	5.5(4)	5.4(2)	18.8(8)	17.5(5)	12.0(3)	24.0(1)	3.5(1)

PROGRAMS	MISSED BOUNDARIES	EXTRA BOUNDARIES	MISSED PLUS EXTRA	# SECONDARY BOUNDARIES
IDS	3	13	16	10
B	5	15	20	3
C	20	5	25	0
D	2	19	21	0
E	10	9	19	0

APPENDIX C

THE 6 PARAMETERS OF REFERENT ALLOPHONES FOR PAIR-WISE PHONE RECOGNITION STUDY

STOP CONSONANTS:

PHONE	CYCLE-FREQ P #1(Hz)	CYCLE-FREQ DISP P #2	DURATION P #3(msec)	TV (NORMALIZED) P #4	MICROSTRUCTURE P #5	ABSAMP P #6
MALE 1						
/g/	1672	720	82.9	614	117	97
/k/	1455	873	70.3	616	168	49
/b/	2296	463	6.0	774	115	56
/p/	2485	322	33.4	787	76	77
/d/	2770	1084	35.1	997	195	94
/t/	3404	1523	101.1	991	169	118
MALE 2						
/g/	1436	428	35.5	468	88	88
/k/	1577	469	154.6	535	57	214
/b/	1562	478	15.7	686	135	31
/p/	2046	544	110.2	707	82	110
/d/	2144	754	60.0	721	108	70
/t/	3218	1544	59.8	1088	246	27
FEMALE 1						
/g/	1872	355	22.5	509	143	112
/k/	1551	260	81.8	489	101	82
/b/	1478	454	17.3	559	81	44
/p/	1911	935	32.8	666	114	37
/d/	2201	551	29.3	563	183	73
/t/	2708	1083	88.7	604	139	48

ALLOPHONE RATIOS

MALE #1

	CYCLE-FREQ	CYCLE-FREQ DISP	DURATION	TV (NORMALIZED)	MICROSTRUCTURE	ABSAMP
A	2.096	1.489	0.408	1.607	1.029	0.973
C	0.868	0.378	0.550	0.824	0.729	0.876
-	1.123	1.784	0.276	1.064	1.035	0.940
N	0.641	0.645	0.558	0.803	0.863	0.770
R	0.759	0.364	0.511	0.800	0.746	0.823
T	1.342	2.091	0.411	1.331	2.138	0.162
W	0.936	1.644	0.397	1.038	1.597	1.129

STANDARD DEVIATIONS

	P #1(Hz)	P #2	P #3(msec)	P #4	P #5	P #6
A	1.064	0.420	0.116	0.312	0.329	0.072
C	0.063	0.037	0.105	0.045	0.016	0.024
-	0.097	0.560	0.033	0.036	0.062	0.115
N	0.027	0.114	0.131	0.060	0.127	0.092
R	0.066	0.108	0.043	0.030	0.025	0.054
T	0.061	0.024	0.280	0.036	1.231	0.061
W	0.036	1.256	0.018	0.066	0.853	0.952

ALLOPHONE RATIOS

MALE #2

	P#1(Hz)	P#2	P#3(msec)	P#4	P#5	P#6
A	1.696	1.387	0.717	1.666	1.383	0.654
C	1.337	1.435	0.076	1.132	1.165	1.075
-	1.068	1.282	0.966	1.032	0.952	1.252
N	0.703	0.655	0.434	0.771	0.856	0.733
R	0.786	0.491	1.003	0.763	0.398	1.561
T	1.065	1.558	0.547	1.120	1.838	0.616
W	0.950	1.105	0.497	0.887	0.975	1.265

STANDARD DEVIATIONS

	P#1	P#2	P#3	P#4	P#5	P#6
A	0.003	1.481	0.956	0.023	0.000	0.510
C	0.007	0.100	0.004	0.002	0.398	0.028
-	0.026	0.307	0.426	0.020	0.168	0.759
N	0.017	0.168	0.127	0.006	0.223	0.171
R	0.024	0.074	0.672	0.012	0.016	2.086
T	0.151	0.789	0.411	0.081	0.669	0.201
W	0.023	0.324	0.274	0.019	0.226	1.195

ALLOPHONE RATIOS

FEMALE #1

	P#1(Hz)	P#2	P#3(msec)	P#4	p#5	P#6
A	1.769	2.693	0.461	1.061	3.193	1.081
C	0.677	0.697	0.378	0.772	0.858	0.471
L	1.041	1.009	0.609	1.057	1.531	1.104
N	0.648	0.907	0.381	0.824	0.523	1.240
R	0.666	1.163	0.752	0.987	0.669	0.851
T	0.877	1.226	0.605	1.095	0.681	0.751
W	0.624	0.925	0.650	0.817	0.602	1.251

STANDARD DEVIATIONS

	P#1	P#2	P#3	P#4	P#5	P#6
A	0.125	0.067	0.009	0.007	0.018	0.070
C	0.000	0.001	0.007	0.003	0.373	0.063
L	0.15[illegible]	0.851	0.157	0.026	0.564	0.143
N	0.173	0.164	0.112	0.035	0.044	0.926
R	0.006	0.375	0.124	0.002	0.097	0.022
T	0.010	0.084	0.413	0.096	0.139	0.004
W	0.009	[illegible]	0.235	0.027	0.103	0.432

THE 6 PARAMETERS OF REFERENT ALLOPHONES FOR PAIR-WISE PHONE RECOGNITION STUDY

FRICATIVES:

MALE #1

PHONE	CYCLE-FREQ	CYCLE-FREQ DISP	DURATION	TV (NORMALIZED)	MICROSTRUCTURE	ABSAMP
	P #1 (Hz)	P #2	P #3(msec)	P #4	P #5	P #6
/f/	4341	2237	184.7	1359	315	38
/v/	3667	2409	45.0	1166	313	63
/θ/	2667	1704	182.0	1079	324	32
/ð/	2035	2069	92.2	920	348	43
/ʃ/	3845	1559	191.2	1111	169	202
/ʒ/	3585	1307	125.7	1081	176	256
/s/	5198	1189	181.4	1372	223	284
/z/	4914	1253	158.9	1333	234	191
MALE #2						
/f/	2646	1113	164.8	913	175	42
/v/	1964	1079	75.1	769	186	28
/θ/	2520	1298	182.0	1011	265	34
/ð/	1776	565	16.5	656	110	76
/ʃ/	3015	864	274.6	938	125	285
/ʒ/	3167	781	127.3	972	130	125
/s/	3735	827	294.3	1221	174	98
/z/	4378	1355	183.4	1236	221	56
FEMALE #1						
/f/	2562	1020	139.7	909	199	32
/v/	1118	472	9.4	373	38	42
/θ/	1390	847	46.4	561	146	59
/ð/	647	651	72.9	364	128	30
/ʃ/	3744	576	262.7	1097	140	159
/ʒ/	2905	1135	118.4	926	160	82
/s/	2690	1411	191.6	973	253	40
/z/	2778	2096	54.6	840	234	83

ALLOPIIONE RATIOS

MALE #1

	P#1(Hz)	P#2	P#3(msec)	P#4	p#5	?#6
W	1.039	1.122	0.985	1.002	1.025	0.899
R	0.917	0.986	0.784	0.958	0.953	0.640
N	0.915	0.959	0.669	0.929	1.021	0.852
T	1.005	0.863	0.640	0.953	0.966	0.857

STANDARD DEVIATIONS

	P#1	P#2	P#3	P#4	P#5	P#6
W	0.029	0.088	0.115	0.015	0.040	0.042
R	0.003	0.024	0.034	0.001	0.013	0.018
N	0.088	0.025	0.069	0.022	0.031	0.036
T	0.133	0.107	0.062	0.048	0.072	0.025

MALE #2

ALLOPIIONE RATIOS

	P#1(Hz)	P#2	P#3(msec)	P#4	p#5	?#6
W	0.981	1. 108	0.896	0.986	1.065	0.981
R	1.007	0.918	0.805	0.968	0.905	1.051
N	0.974	0.929	0.831	0.987	0.923	0.969
T	1.017	1.015	0.952	1.000	1.202	0.743

STANDARD DEVIATIONS

	P#1	P#2	P#3	P#4	P#5	P#6
W	0.036	0.236	0.098	0.027	0.131	0.110
R	0.027	0.016	0.037	0.003	0.015	0.073
N	0.031	0.072	0.167	0.007	0.074	0.090
T	0.083	0.146	0.095	0.006	0.036	0.000

FEMALE #1

ALLOPHONE RATIOS

	P#1(Hz)	P#2	P#3(msec)	P#4	p#5	?#6
W	0.964	0.949	0.836	0.926	0.937	1.059
R	1.063	1.433	1.025	1.382	0.829	1.508
N	1.264	0.975	1.141	1. 121	1.084	2.191
T	0.939	1.124	0.622	1.023	1.341	0.592

STA''ARD DEVIATIONS

	P#1	P#2	P#3	P#4	P#5	P#6
W	0.268	0.028	0.085	0.080	0.027	0.042
R	0.083	0.660	0.061	0.075	0.009	1.325
N	0.370	0.099	0.616	0.057	0.163	6.831
T	0.068	0.071	0.044	0.030	0.059	0.064

STOP CONSONANTS

PAIR-WISE PHONE COMPARISON TEST RESULTS

MALE #1

	G	K	B	P	D	T
G	15	1	1	1	2	
K	2	9	1	1	1	
B			9			
P		1		3		
D	1	1	1	1	19	3
T	2	3	1	2	1	17

MALE #2

	G	K	B	P	D	T
G	14	1				
K		15				
B			10	1	1	1
P		2		11		3
D				3	8	1
T					2	13

FEMALE #1

	G	K	B	P	D	T
G	15		1	1		3
K	2	9	1	1		2
B	1		9	2		1
P			1	7	1	
D			2	2	10	2
T			1	2	2	7

FRICATIVES

PAIR-WISE PHONE COMPARISON TEST RESULTS

MALE #1

	F	V	θ	ð	ʃ	ʒ	S	Z
F	21		3					
V		17						
θ		2	18					
ð		2		20				
ʃ					21	2		
ʒ				1		12		
S							26	
Z							2	21

MALE #2

	F	V	θ	ð	ʃ	ʒ	S	Z
F	18	1	2	1				
V		13	2	1				
	1		10	1				
				7				
				1	21	1		
				1		13		
S							26	2
Z				1				24

FEMALE #1

	F	V	θ	ð	ʃ	ʒ	S	Z
F	19				3			1
V		1						
	2	1	6	2				
		1		8				
		1		1	15			
		1		1	1	13		1
S		1		1	1		28	2
Z		1		1	1			24

BIBLIOGRAPHY

[A1[Atal, B.S. and S.L. Hanauer, Speech analysis and synthesis by linear prediction ot the speech wave, JASA 55: 637-655,1974.

[B1] Baker, J.M, J.K. Baker, and J.Y. Lettvin, More visible speech, JASA 52: 183 (A), 1972.

[B2] Baker, J.M., R. Ramsey, M. Miller, J.K. Baker, and C. Cooper, Comparative visual displays of time and frequency domain information in connected speech, JASA 55 (no 2): (A), 1974.

[B3] Baker, J.M. and R.T. Schumacher, Computer study of "jitter" in violin and cello tones, paper presented at "A Topical Conference on the Teaching of Acoustics and the Physics of Sound and Music", April 5-6,1974, Univ. of Iowa, Iowa City, Iowa.

[B4] Baker, J.M., A new time-domain analysis of fricatives and stop consonants, Proc. of IEEE Symposium on Speech Recognition, Pittsburgh, Pa., 1974.

[B5[__, Time-domain acoustic characteristics of allophones and phonological phenomena, JASA 55: (A), Supplement, Spring, 1974.

[B6[__, Time-domain analysis and segmentation of connected speech, Proc. of the Speech Communication Seminar, Stockholm, 1974.

[B7[__, Automatic time-domain techniques for segmentation of connected speech, JASA 56: (A), Supplement, Fall, 1974.

[B8[__, New time-domain analysis for complex animal vocalizations, JASA 56: (A), Supplement, Fall, 1974.

[B9[Boomsliter, P.C. and W. Creel, Research potentials in auditory characteristics of violin tone, JASA 51: 1984, 1972.

[C1[Chandra, S., Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis, IEEE Trans. Acoust., Speech, Signal Processing, ASSP-22 (no.6): 403-415, 1974.

[C2[Chang, S.H., G.E. Pihl, and J. Wiren, The intervalgram as a visual representation of speech sounds, JASA 23 (no. 6): 675-679, 1951.

[C3[Chang, S.H., G. Pihl, and M.W. Essigmann, Representation of speech sounds and some of their statistical properties, Proc. I.R.E. 39: 147-153, 1951.

[C4[Chang, S.H., Two schemes of speech compression system, JASA 28: 565-572, 1956.

[D1[Davis, H. Peripheral coding of auditory information, in *Sensory Communication* (W. Rosenblith, Ed.), MIT Press,119-141,1961.

[D2[Davis, K.H., R. Biddulph, and S. Balashek, Automatic recognition of spoken digits, JASA 32: 1450-1455, 1960.

[F1] Frishkopf, L. and M. Goldstein, Responses to acoustic stimuli from the eighth nerve of the bullfrog, JASA 35: 1219-1228, 1963.

[G1] Galambos, R. and H. Davis, The response of single auditory nerve fibers to acoustic stimulation, J. Neurophysiol. 69: 58, 1943.

[G2] Gerstman, L.J., Classification of self-normalized vowels, IEEE Trans. Audio Electroacoust., AU-16 (no. 1): 78-80,1968.

[I1] Ito, M.R., Investigation of time domain measurements for analysis of speech, Ph.D. Thesis, Univ. of British Columbia, 1971.

[K1] Kiang, N.Y.S., *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, MIT Research Monograph, No. 35, 1965.

[K2] __, and E.C. Moxon, Tails of tuning curves of auditory nerve fibers, presented at the 85th meeting of the Acoustical Society of America, April 1, 1973.

[K3] Konishi, M., Time resolution by single auditory neurones in birds, Nature 222 (no. 5193): 566-567.

[K4] __, Comparative neurophysiological studies of hearing and vocalizations in songbirds, Z.vergl.Physiologie 66: 257-272.

[L1] Licklider, J.C.R. and I. Pollock, Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech, JASA 20: 42-51, 1948.

[L2] Licklider, J.C.R., The intelligibility of amplitude-dichotomized time-quantized speech waves, JASA 22: 820-823, 1950.

[M1] Makhoul, J.I. and J.J. Wolf, Linear prediction and the spectral analysis of speech, Bolt, Beranek, and Newman. Inc., Cambridge, Mass., Report #2304,1972.

[M2] Markel, J.D., The Prony method and its applications to speech analysis, JASA 49: 105 (A), 1971.

[M3] Morris, L.R., The role of zero crossings in speech recognition and processing, Ph.D. Thesis, Univ. of London, 1970.

[M4] Munson, W.A., and H.C. Montgomery, A speech analyzer and synthesizer, JASA 22: 678 (A), 1950.

[P1] Peterson, E., Frequency detection and speech formants, JASA 23: 668-674, 1951.

[P2] Pollock, I., Detection and relative discrimination of auditory "jitter", JASA 43: 308-315, 1968.

[R1] Reddy, D.R., Segmentation of speech sounds, JASA 40 (no.2):, 307,1966.

[R2] Rose, J.E., J. Brugge, D. Anderson, and J. Hind, Phase-locked responses to low-frequency tones in single auditory nerve fibers of the squirrel monkey, J. Neurophysiol. 30 (no. 4): 767-793, 1967.

[S1] Sakai, T. and S. Inoue, An analyzing equipment for the zero crossing interval and its applications to speech analyses, J. Inst. Elect. Commun. Engrs. Japan 39: 404-409, 1956, (in Japanese), English abstraction in Phys. Abst. 60: 110-1157,1957.

[S2] Schatz, C.D., The role of context in the perception of stops, Language 5: 47-56, 1954.

[S3] Shoup, J., The phonemic interpretation of acoustic phonetic data, Ph.D. Thesis, Univ. of Michigan, 1964.

[S4] Stevens, K.N. and A. House, Perturbation of vowel articulation by consonant context: an acoustic study, J. of Speech and Hearing Res. 6: 111-128, 1963.

[S5] Stevens, K.N. and D.H. Klatt, Role of formant transitions in the voice-voiceless distinction for stops, JASA 55 (no. 3): 653-659, 1974.

[T1] Thorpe, W.H., Antiphonal singing in birds as evidence for avian auditory reaction time, Nature 197: 774-776, 1963.

[T2] Thorpe, W.H. and M.E.Y. North, Origin and significance of the power of vocal imitation: with special refernce to the antiphonal singing of birds, Nature 208: 219-222, 1965.